

Chapter 2

Analytics of Risk and Challenge

The formulation of a problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill. To raise new questions, new possibilities, to regard old problems from a new angle, requires creative imagination and marks real advance in science

Albert Einstein and Leopold Infeld (1938) [2]

Abstract As emphasized several times in the previous chapter, CRT is about analyzing risk and designing deliberate challenges. Whether we are deliberately challenging the effectiveness of a strategic plan or the scalability of an optimization or big-data mining algorithm, the concept of a challenge has the same fundamental characteristics. The purpose of this chapter is to develop a disciplinary approach to structure and model the analysis of risk and the concept of a challenge. This structure can assist an automated system to risk assess and challenge a human or a computer autonomously, and to teach the concept of challenge in a disciplinary manner to humans. What is risk? How to analyze risk and how to “think” risk? What is a challenge? What do we mean by deliberate? How do we design and model the concept of a challenge deliberately? How do we systematically design a challenge on which both humans and computers to operate? This chapter will address these questions by establishing a unifying theory that defines and models systems, uncertainty, ability, skill, capacity, competency, performance, capability, and our ultimate aim, risk and challenge.

2.1 Precautions

This chapter will revisit many basic concepts that may seem already known to many readers. Nevertheless, a formal definition of each of these concepts will be provided. Some of the definitions will be obvious, some may deviate from daily uses of the concept, and some may even contradict our present understanding of the concept. This is why defining these basic concepts is essential.

The discussion of many concepts in this chapter intersects with other disciplines, including those of the behavioral and educational sciences and organizational psychology. In fact, psychology literature is rich in dealing with these concepts, with many articles published on each of the many concepts that will be discussed here.

A CRT exercise may include a behavioral psychologist to perform a behavioral assessment of the blue team. It may use an organizational psychologist to understand the culture of the blue organization or it may include a cognitive psychologist to advise on task designs with specific cognitive-load characteristics to overload blue's thinking. Our discussion in this chapter does not aim to discuss these roles and the science needed to perform each of them. A psychologist in any of these roles is another team member of the CRT exercise, bringing their own expertise to the CRT exercise. Psychology literature examines each of these roles and concepts underpinning them with more depth than the discussion here.

The discussion in this chapter does not aim to reproduce the psychology literature, nor does it aim to introduce a new psychological theory. The main aim is to design a model of a challenge that we can use in a computational environment. This model will be used to analyze an algorithm, a machine, a human or an organization. The discussion will offer simple and structured behavioral models that can be used by non-psychologists. These models are simple when compared to the great amount of literature available on the concepts, and the complexity involved in understanding human psychology. However, the models are reliable because whether we use pencil and paper or computers to red team, and whether we use them for small or large-scale problems, they will produce results that can be traced to causes and evidence.

To bring the different pieces of a model of challenge together successfully, the discussion will intersect with a number of fields, including psychology, education, risk management, system theory, and computational sciences. Structuring these concepts is a daunting task. First, science by nature offers a thesis and antithesis. The reader may find scientific articles with different definitions that contradict each other. In places, the treatment of the topic will certainly contradict some of this science. Second, most of these concepts are also used in our daily language; therefore, a first encounter with a definition for any of these concepts that does not comply with one of our daily uses may create unease for the reader.

Nevertheless, given that one of the aims is to structure these concepts so that we are able to compute them, we must define them clearly in an unambiguous manner. Such unambiguous definitions will eliminate confusion in the reader's mind while reading this book, even if the definitions themselves are not universally accepted.

2.2 Risk Analytics

We define risk analytics as follows:

Definition 2.1. Risk analytics is the process of transforming data and requirements into actions using risk thinking and a disciplined risk methodology to understand historical situations, anticipate and predict futures, select appropriate courses of actions for an organization to implement, and/or determining novel ways for an organization to operate.

The above encompassing definition covers the roles and benefits of risk analytics within an organization. To illustrate risk analytics as a process, Fig. 2.1 presents six standard steps. These steps are very similar to those followed in any type of decision making situation. However, risk analytics emphasizes that the overall decision making process is guided with, and centered on, the concept of risk.

The first step is related to intelligence gathering and reconnaissance operations, that is, the process of collecting data and targeted evidences to support the decision making process. In the military and security domains, intelligence and reconnaissance are two classic functions that drive any operation. Similarly, in businesses, the field of business intelligence has witnessed large interest to provide the data required to steer the decision making process. In government, evidence-based policy is normally the terminology used to stress the need for having the right data to shape policy development.

Intelligence does not only react to the needs of the organization, but also provides a proactive capability to shape and drive organizational needs. As data gets



Fig. 2.1 Risk analytics steps

collected, the organization continuously assesses the situation, the associated risks, and the threats that may exist in the environment. Most of these terminologies, such as risk and threats, will be explained in more details in the rest of this chapter. For the time being, we can rely on common knowledge in understanding these terminologies to follow the current discussion on the risk analytics process.

When the organization identifies a specific type of threat or a possible negative or positive impact on organizational objectives, a need arises to analyze this situation and formulate alternatives. Response analysis is the process of formulating and assessing responses. Consequence analysis then projects some of the selected responses onto future states to assess the longer term impact of these responses on organizational objectives.

A suitable response is then selected. The response design step transforms the selected response into suitable actions that can be executed. For example, one important aspect of response design is how the selected response will be framed to others. The organization may decide to fire people. Will the organization present this response as a direct consequence of drop in sales, as a restructure of operations to improve productivity, or as a step towards renewing the organization. Framing the response is a very critical skill that can dramatically impact the effectiveness of the response in achieving the intended impact.

When risk analytics relies on designing challenges as the tool to react to threats, the process gets more targeted, where the threat actor becomes the focal point of the analysis. In other words, intentional actions become more paramount in the analysis, as well as the response.

2.2.1 Intentional Actions

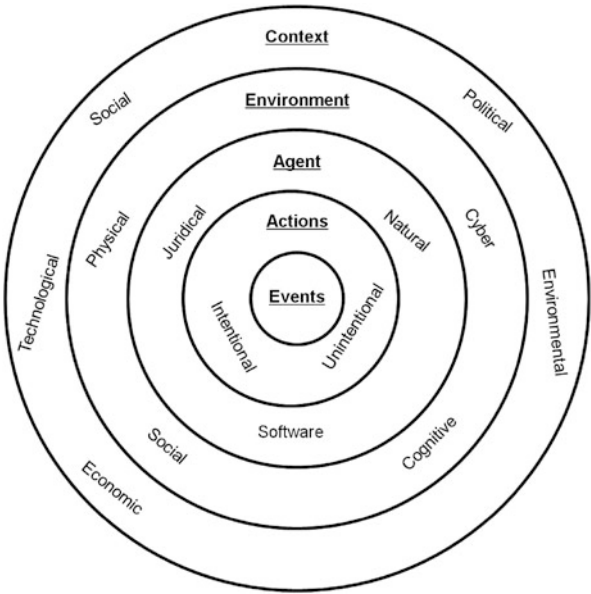
CRT is designed to challenge an entity. The success of our ability to challenge this entity must be reflected in its performance. In CRT, this entity can be anything from a human to a machine, from a company to a country, and from a technology to ideas and beliefs. Regardless of what the entity is, it needs to have an owner. We will call the owner a person or an agent.

We will follow a legal definition of a “legal person,” or a person for short. A person can be natural, as a human, or a juridical, as a corporation. We will reserve the word “agent” to mean both a person and software that performs some tasks by producing actions. We will use the word “entity” to refer to agents that think/compute and act or objects that do not think or act. If our discussion is limited to a software agent, we will explicitly refer to it as a “software agent.”

While a person and an agent are systems by definition, as will be discussed in this section, the word “system” will be used to emphasize the structure over the identity, and the words “person” or “agent” will be used to emphasize identity over structure.

Whatever the type of an agent, we will consider an agent as a living organism: it continuously produces actions in the environment. Even if the agent stays still, staying still is an action. When a human or a computer program goes to a sleep, this

Fig. 2.2 Understanding agents’ actions and their relationships to the environment



is an action in its own right. Therefore, an agent produces actions continuously in the environment; each action will produce outcomes.

An agent (see Fig. 2.2) lives within four different generic environments that we will call them the four domains. These are social, cognitive, cyber, and physical domains (SC2PD). These environments are surrounded with many different contexts, including the PESTE contexts. An agent lives within these contexts and environments, and impact them by generating actions which create events that influence the context, environment and the agent.

CRT is not a context that concerns reflex, unintentional or ad-hoc actions. A red team is established for a purpose, and with an understanding of who the blue team is. Therefore, in the context of CRT, we will focus on intentional actions.

Definition 2.2. An intentional act is the production of an action by an agent to fulfil the agent’s goals.

Therefore, these intentional actions are not produced in vacuum; they are produced to achieve an agent’s goal. This does not necessarily mean that an action successfully achieves the goal of the agent. At the time the action was produced, the agent’s intent was to achieve an agent’s goal, irrespective of whether this action was actually successful in achieving this goal. This produces deviations between the actual outcome of actions and the intended outcomes by the agent. When these deviations are sensed by the agent, they act as a feedback signal for the agent to adjust its set of actions accordingly.

2.2.2 *Objectives and Goals*

A properly designed intentional action needs to consider the outcomes the agent intended to achieve the fulfilment of the agent's objectives, goals and the uncertainty surrounding the achievement of these outcomes. This begs the question of what these concepts mean.

Definition 2.3. An objective is an approximately measurable phenomenon with a direction of increase or decrease.

The phenomenon can be the agent's state. For example, when an affective state such as happiness or a physical state such as monetary richness become the subject of an objective, we would usually have a metric to measure this state. In the case of the affective state of happiness, we may not have a direct manner by which to measure the state itself, but we can use a set of indicators. These indicators are blended (fused) to provide a measurement of the degree of happiness. We would then either attempt to increase (maximize) or decrease (minimize) the degree of happiness.

In CRT, the objectives of both teams are somehow interdependent because the agent's states are interdependent on each other. For example, the red team's affective state of happiness may be negatively influenced by the blue team's state of richness (as in the simple case of human jealousy); thus, a decrease in blue's richness generates an increase in red's happiness. In this case, the red team may have an objective of minimizing the richness of the blue team to maximize its own happiness. If the teams' objectives are independent of each other, they should act independently; therefore, there is no need for the CRT exercise in the first place.

If red and blue objectives are positively correlated,¹ they can optimize their objectives either by continuing to act independently, or by taking an opportunity that might arise to optimize their objectives by acting cooperatively. In this case, the objective of the CRT exercise is to explore novel opportunities for collaboration.

However, in most cases, CRT exists for competitive situations.² In this case, a blue-red competition can only exist if blue and red have conflicting objectives. Conflicting objectives can take two forms. In the first form, the objectives themselves are in direct conflict with each other. For example, in a situation of war, blue wishes to win at the cost of red losing, and vice versa.

In the second form, the objectives may not be in obvious conflict, but limited resources place them in conflict. For example, there are two departments in a company, one is responsible for research and development (R&D) and the other is responsible for the core-business production line (the production department).

¹Two objectives are said to be positively correlated if an improvement in one is accompanied with an improvement in the other and vice versa.

²Even when we discuss CRT for cooperative situations, we use competition as the way to achieve cooperation. For example, by challenging the student's mind with stimulating ideas, the student becomes more engaged, and pays more attention to and cooperates with the teacher.

The R&D department’s objective is to maximize the innovation of its design of the next generation of products. However, the production department has objectives such as maximizing production efficiency and product quality. While the objectives of both departments are almost independent because the output of each department aims at different time-scales, the fact that their budget comes from a common pool can put them in direct conflict. Here, the conflict is that as one department attempts to draw resources to achieve its objectives, it is depleting and competing with the resources available to the other department for achieving the other department’s objectives.

The same forms of conflicting objectives occurs in CRT. For example, a CRT exercise to evaluate the security system of a company would place both teams in direct conflict. The blue team’s objective is to maximize the protection of the security system, while the red team’s objective is to maximize the probability of breaking into the security system. This arm race is a core characteristic of the CRT exercises.

This discussion demonstrates the importance of mapping out and understanding the objective space for both red and blue in a CRT exercise. Figure 2.3 presents a conceptual objective space for both red and blue. A solid arrow/line between two objectives indicates positive influence. For example, if *or2* in the figure represents richness and *or4* represents happiness, the solid arrow from *or2* to *or4* indicates that as richness increases, happiness also increases. A line, instead of an arrow, indicates influence in both directions.

It is critical to analyze this objective space in a CRT exercise because of the interdependency between objectives. For example, we can see that *or7* for red has a

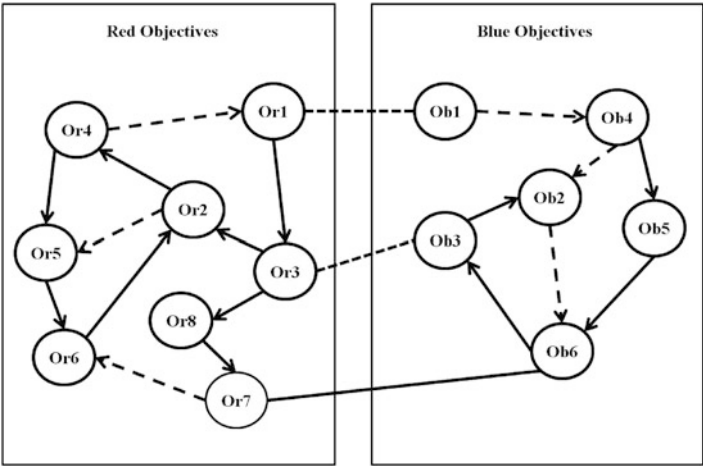


Fig. 2.3 Blue and red objective spaces and their correlations. A solid arrow/line indicates positive correlation; a dotted arrow/line indicates negative correlation

positive relationship with *ob6* for blue. That is, it is beneficial for both blue and red to cooperate to maximize these objectives.

However, this conclusion is superficial. We need to understand the complex interteam and intrateam interactions in the objective space.

For blue, *ob6* positively influences *ob3*, while an improvement in *ob3* will improve *ob2*, which will negatively influence *ob6*. This generates a negative cycle with blue's objective space. For example, improving education intake and quality would improve health, but improving health would increase the age of retirement, degrading job market, which then negatively influences education. Similarly, in a network-security scenario, creating a stronger security system through multiple biometric authentication protocols would increase system protection, but increasing system protection would reduce the usability of the system (customers need to spend more time to authenticate), which may increase customer dissatisfaction. These examples demonstrate the internal conflict that can exist within the interteam objective space.

This creates an internal conflict within blue objectives. Blue would then need to establish its own trade-offs. In the meantime, red does not have the same internal conflict. *or7* negatively influences *or6*, which positively influences *or2*, which positively influences *or4*, which negatively influences *or1*, which positively influences *or3*. That is, *or7* positively influences *or3* (if we multiply all signs on the path, we obtain a positive sign). We notice that there is a conflict between *or4* and *or1*, but this conflict does not impact the interdependency between red's external objectives.

If we examine the intrateam interaction, we see that *ob6* for blue positively influences *ob3* for blue, which negatively influences *or3* for red. Therefore, blue has the following two problems:

1. Blue has a negative feedback cycle internally: $ob3 - ob2 - ob6 - ob3$. Red can influence this negative feedback cycle as red's *or7* objective interacts positively with blue's *ob6* objective. Thus, red can influence blue's decision made on any internal level of trade-off.
2. Red's *or3* and *or7* objectives reinforce each other. In the meantime, red's *or3* objective is in conflict with blue's *ob3* objective. As red improves its own *or3* objective, blue's *ob3* objective deteriorates.

Once these objectives become known, each team attends to design plans to achieve their objectives. To monitor progress toward the objectives, goals are defined.

Definition 2.4. A goal is a planned objective.

Based on the agent's assessment of what is possible and what is not, the agent can establish an "aspiration level" for each objective. This process of planning and designing aspiration levels transforms each objective, where the agent wishes to optimize the objective, to goals, where the agent wishes to reach the way-point indicated by the aspiration level.

In classical optimization, the problem the agent wishes to optimize can be formulated as follows:

$$\begin{aligned} &\downarrow f(x) \\ \text{S.T. } &x \in \Phi(x) \end{aligned}$$

where, $f(x)$ is the objective the agent wishes to optimize (minimize in this case), x is the decision variable(s), the alternatives or courses of action from which the agent needs to choose, and $\Phi(x)$ is the feasible space of alternatives. Every solution belonging to the feasible space $\Phi(x)$ satisfies all constraints in the problem. We use \downarrow to denote minimization, \uparrow to denote maximization, and “S.T.” as a shorthand for “subject to the following constraints or conditions.”

For an agent to optimize one of its objectives, it needs to form a plan, or a series of actions to make this optimization work. The agent’s plan is designed after careful assessment of what is possible and what is not, or what we will term “constraints.” Once planning is complete, the agent becomes more aware of the environment, as well as what it can achieve and what it cannot. In this case, the objective is transformed into a goal and the formulation above can be re-expressed as is presented in the following equation.

$$\begin{aligned} &\downarrow d^- + d^+ \\ \text{S.T. } & \\ &f(x) + d^- + d^+ = T; \\ &x \in \Phi(x) \end{aligned}$$

where T is the target or aspiration level of the goal, d^- is the underachievement of a goal, and d^+ is the overachievement of a goal. In this formulation, $f(x) + d^- + d^+ = T$ is termed a “soft constraint”, while $x \in \Phi(x)$ is termed a “hard constraint”. A feasible solution can violate a soft constraint with a cost, but it can’t violate a hard constraint. The objective function can take many forms including the minimization of underachievement alone, overachievement alone, or a weighted sum of both.

Figure 2.4 presents a pictorial diagram to emphasize the difference between interteam and intrateam conflicting objectives. As we discussed above, each team has its own internal conflicting objectives. Each team needs to decide on the level of trade-off to compromise the optimization of these internal conflicting objectives. In the meantime, blue-red interaction has its own conflicting objectives. A level of trade-off is still necessary, as both red and blue need to compromise. Therefore, Both interteam and intrateam conflicting objectives generate two different decision-science problems that need to be solved. However, the tools used to solve interteam conflicting objectives significantly differ from those used to solve intrateam conflicting objectives because of the following three reasons:

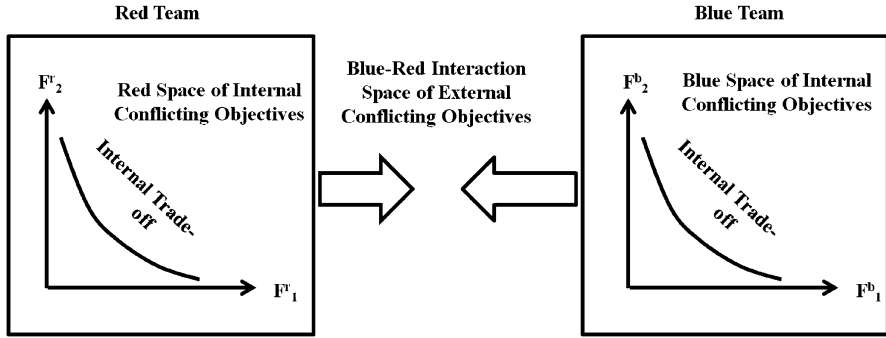


Fig. 2.4 Differentiating interteam and intrateam conflicting objectives

1. The first difference lies in who owns the trade-off. For the interteam conflicting objectives, each team owns their problems and therefore can decide on the level of trade-off they wish to achieve. In the intrateam conflicting objectives, the trade-off is owned by both teams together. The issue of ownership is core when selecting an appropriate technique to solve these problems because it defines the level of control of a team on implementing a proposed solution. One would expect that red and blue could exercise more control internally than externally.³ The implication here is an internal decision made by one team will be easier to implement than an external decision.
2. The second difference lies in the nature of the trade-off. In the intrateam conflicting objective space, the trade-off is not usually a one-off decision; it needs to be negotiated and be determined by both teams together. As blue makes a decision, red responds, and as red makes a decision, blue responds. Therefore, the trade-off in the intrateam conflicting objective space is more dynamic than in the interteam conflicting objective space.
3. The third difference lies in the nature of uncertainty and information availability in the intrateam and interteam conflicting objective space. In an interteam situation, external uncertainty is almost uncontrollable. The system attempts to decide on its actions to manage the risk of these external uncertainties. In the intrateam situation, uncertainty is dynamic. As the two teams interact, their actions can shape the uncertainty space. This discussion point will be revisited in Sect. 4.1.

By now, we should ask whether the division between internal conflicting objectives and external conflicting objectives is meaningful. In fact, this division largely depends on where we draw “system boundaries.” In the following section,

³How to deal with the situation when one of the teams has more control externally than internally is outside the scope of this book.

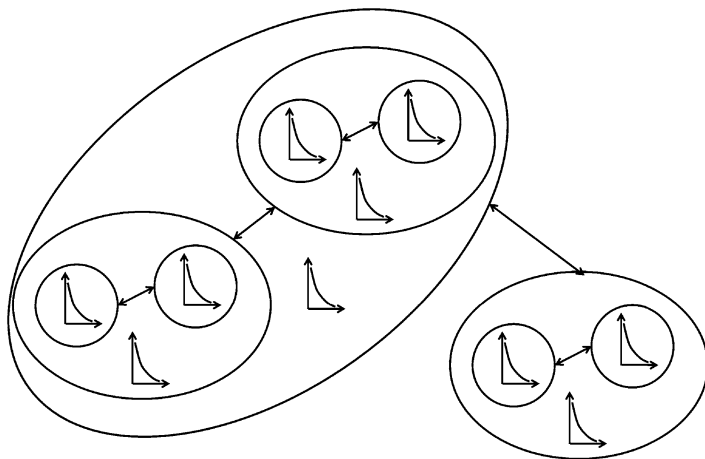


Fig. 2.5 The nested nature of red teaming

a “system” is defined. However, to illustrate that the division drawn between intrateam and interteam is artificial, and that CRT is not simply an exercise between “us and them,” Fig. 2.5 conceptually depicts the nested nature of CRT.

Figure 2.5 demonstrates that for whatever system we red team, within this system, we can have another CRT exercise. For example, organization *A* may red team its strategy against organization *B*. According to the previous discussion, *A* owns its objectives and decisions. However, within *A*, there are departments with conflicting objectives. Each department can conduct its own CRT exercise and perceive the rest of the departments as external teams. Within the same department, different sections may have conflicting objectives and they may apply CRT to evaluate their own strategies. Within a section, people may have conflicting objectives and a person may use CRT to evaluate their own plans. In short, CRT is not an exercise between one country and another alone; as discussed in this book, it is an exercise that can be used for individuals, organizations, countries, and even algorithms and machines.

This demonstrates that the CRT exercise is also nested by definition. When red and blue are two organizations, each team can be divided into smaller teams. There may be a red team for the internal budget, another for the internal market strategy, then the two teams may form a red team for the organization given that they are in a better position to understand the internal and external objectives and uncertainties.

Some people argue that this nested view of CRT is not desirable because CRT is perceived as an exercise with the enemy; so how can we red team inside the organization? Reasoning demonstrates that within the organization, senior management can resolve conflict, but if two countries fight, there is no equivalent concept to senior management in organizations. Therefore, there is a fundamental difference between CRT exercises conducted between organizations, and those conducted within an organization.

This argument is flawed in two aspects. First, it reflects the limited view that CRT is a military or national-security exercise. Limiting the concept of CRT to these domains will harm these domains because the constrained context, while important, limits the possibilities for CRT to grow as a science.

The second reason the argument is flawed is that the concept of senior management exists in every problem. Senior management is not an external counseling service or a legal authority. Members of senior management come from different portfolios in an organization. Even for matters related to military or national security, different countries are members of a larger international organization such as the United Nations. This does not eliminate the need for CRT on a country level, a state level, a department level, an organization level, a technological level, or even an algorithmic level. CRT is a nested exercise simply because conflict in objectives is a nested concept. The fact that larger objectives are comprised of smaller objectives can create conflict itself, and as each person is responsible for a different portfolio within an organization, CRT on one level is comprised of CRT exercises on sublevels.

2.2.3 *Systems*

As discussed, the primary reason that red and blue are in conflict is that the objectives of the blue system are in conflict with the objectives of the red system. In a CRT exercise, it is critical to consider both the red and blue teams as a system. For red, blue is a system for which red attempts to dysfunction by counteracting its objectives. The same is true for blue, red is a system that is attempting to dysfunction blue because red's objectives are in conflict with blue's objectives. We use the word "dysfunction" since interference with a system's objectives with the aim of acting against the benefits of the system is a possible cause for dysfunction. This dysfunction can take the form of simply influencing the objectives of one team to change, or in more dramatic situations, of damaging the components of the system.

Classically, a system is perceived as a group of components or entities interacting for a purpose. This definition is too basic here, and does not adequately service our analysis. Therefore, a system is defined here as follows.

Definition 2.5. A system is a set of entities: each has a capacity to receive inputs, perform tasks, generate effects, and complement the other toward achieving goals defined by a common purpose.

Definition 2.6. An effect is a measurable outcome generated by an action or caused by a change in a system state.

The above definition of "system" can be considered an elaborate definition of the classical definition of a system. However, this further level of detail is necessary. It makes it clearer to an analyst that when they define a system (such as the red or blue system), they must map out the entities; the inputs to each entity; the task each

entity is performing (reflecting the purpose of this entity or subsystem); the effects that each entity generates; and how these entities and their objectives depend on each other and come together to achieve the overall purpose of the system.

The definition for “effect” clarifies that given actions are produced continuously, effects are also generated continuously. Every action produces many outcomes. An effect is a measurable outcome within the context of CRT. If the outcome is not measurable, it cannot be considered within a CRT exercise before it becomes measurable (either directly or indirectly through a set of indicators); otherwise the exercise will become an ad-hoc activity.

If we want to discuss change in happiness as an effect, we need to know how to measure happiness. Alternatively, we need to find indicators that collectively indicate happiness so we can measure these indicators. If we cannot measure directly or indirectly, we cannot manage, we cannot engineer, we cannot define a reward or penalty, and we simply cannot influence or control.

The definition of “effect” also emphasizes that effects can be produced without actions. For example, aging is an effect of time. Even if we put the human on a bed in a coma, the body will continue to age and decay.⁴ These changes in the state of the system are naturally occurring without actions per se.

The definitions of system and effects used above are particularly useful for a red teamer because they create knobs for engaging with the system to steer it and influence it in a more clear manner. Knowing how the entities interact and the resultant effects provides us with an idea of which entities are more important than others, and which are more controllable than others. Once we define the key entities we wish to control, we can examine how to control them and the desired changes in the effects. However, given that each of these entities is a system, we can continue to deconstruct the problem and locate more control points.

The second group of knobs is the inputs, the tasks an entity is performing, and the effects an entity generates. Chapter 4 will present a more elaborate discussion on this issue. Understanding these knobs facilitates the task of the red teamers.

Components comprising a system are in their own right, a system. An aircraft is a system, as it consists of the mechanical, software, fuel and human components, without which it cannot fulfil its purpose. The purpose of an aircraft is to fly. This is actually an assumption for which we should pause and consider in depth.

Definition 2.7. The purpose of a system is the reason for being from the perspective of an external observer.

While the components are internal to the system, the purpose is always in the eyes of the beholder. The purpose of a system is an external judgment that is made by an external stakeholder or observer. The purpose is defined by an external entity, which can also be the owner of the system. Therefore, the same system can have multiple

⁴One can consider this concept on a philosophical level as actions produced by the environment that cause decay to occur, but we will avoid this level of interpretation in this book because it can create unmanageable analysis.

purposes. For an airline, an aircraft's purpose is to make money through flying. For the post office, an aircraft's purpose is to deliver the mail. For a business passenger, an aircraft's purpose is to provide transportation to attend business meetings. For a world traveler, an aircraft's purpose is to provide transportation to travel from place to place for enjoyment.

The different views on the purpose of an aircraft by different external stakeholders in the community may generate conflicting objectives. Making more profit from an airline perspective can create conflict with a passenger who wishes to minimize the cost of travel as much as possible. A longer route at an optimal altitude may minimize fuel costs for the airline as compared to a shorter route at an inefficient altitude, which burns more fuel. However, for the business passenger, a longer route may entail late arrival at the destination.

For an airline company, the board will define the purpose of the company. One can perceive the board as an external entity, which in reality it is because it represents the interface between the stakeholders of the company and the company itself. The chief executive officer (CEO) sits on the board as an ex-officio and reports to the board. Through the CEO, the purpose is translated into internal objectives, which are then transformed into goals, key performance indicators, and plans.

While the aircraft's purpose for one person is for them to be able to fly, for another, it might be a symbol of power and wealth—imagine having an aircraft in your backyard that you do not intend to use. You only have it on display to show your neighbors how wealthy you are.

In the latter case, it does not matter whether we run out of fuel since the purpose of this aircraft is to symbolize power and wealth, not to fly. It does not even matter if the crew does not arrive or the control software system is not working. These elements are not critical for the purpose.

Therefore, there is a tight coupling between the purpose of a system, and which elements of an aircraft are deemed important for that purpose. Elements contributing to different purposes can overlap. However, all elements of an aircraft may exist, but not all of them are critical elements for the aircraft (the system) to fulfil its purpose. Therefore, what defines the “critical elements” in a system can be different from one observer to another, and from one stakeholder to another.

Definition 2.8. An element or component in a system is termed “critical” if the removal of, or cause of damage to, this element or component would significantly degrade the ability of the system to achieve its objective, goal, or purpose.⁵

For example, the heart is a critical element in the human body because if it is attacked, the human body defining the system in this context will find it difficult to achieve its objectives and its purpose of functioning efficiently and living, respectively.

⁵Most of the definitions used for critical elements, hazards, threats, and risks in this book are compatible with ISO3100 [8], but sometimes get slightly changed to fit the context of this book.

In the example of the aircraft in the backyard as a symbol of power, the critical element of the aircraft is that it has all its exterior body parts, including the wheels. Scratches in the paintwork may not affect its ability to fly, but would certainly affect its appearance as a symbol of power. The engine is no longer a critical component; if it is not working, the appearance is not impacted.

It is clear that what makes a component in the system a critical element is its contribution to the capacity of the system in achieving its purpose. However, neither this capacity nor the objectives are deterministic; they are impacted by both internal and external uncertainties.

2.2.4 Uncertainty and Risk

A properly designed action must consider the outcomes the agent intended to achieve at the time the action was formed to fulfil the agent's objectives or goals, as well as the uncertainty surrounding the achievement of these outcomes. So far, we have discussed objectives and goals. However, the perceived outcomes are the agent's expectation of an action's impact on objectives given the uncertainty of that impact. Many factors come into play in determining this uncertainty, from the personality traits of the agent to the agent's sensorial abilities, availability and access to information for the agent, and the complexity of the situation the agent faces.

Every action must be evaluated through its effects and the impact of these effects on both red's and blue's objectives. These effects need to be designed systematically and consider the uncertainty in the environment. Therefore, in CRT, the concept of risk is paramount.

From an agent's perspective, Fig. 2.6 depicts a basic form of the decision-making cycle an agent undergoes. The agent relies on its sensors to perceive uncertainty in the environment. The agent has a set of feasible actions it wishes to evaluate for the particular context in which it is attempting to make a decision. Together with the agent's objectives, the agent needs to make a judgment on how these uncertainties impact the agent's objectives for each possible action the agent needs to evaluate.

The agent selects a possible action to execute based on the agent's assessment of the impact of uncertainty on objectives if this action is executed. This assessment is also influenced by the agent's risk personality traits and experience. The agent's personality towards risk gets influenced by the agent's perception of uncertainty and the feedback received from the environment; together, they can reshape the agent's attitude to risk.

For example, the manner a message gets framed and presented to an agent influences the agent's perception of the level of uncertainty in the environment. Consider for example the difference between "this person is trustworthy" and "to my knowledge, this person is trustworthy". The second statement can be perceived to carry more uncertainty than the first, despite that we understand that whatever statement someone is making, it is based on the person's level of knowledge.

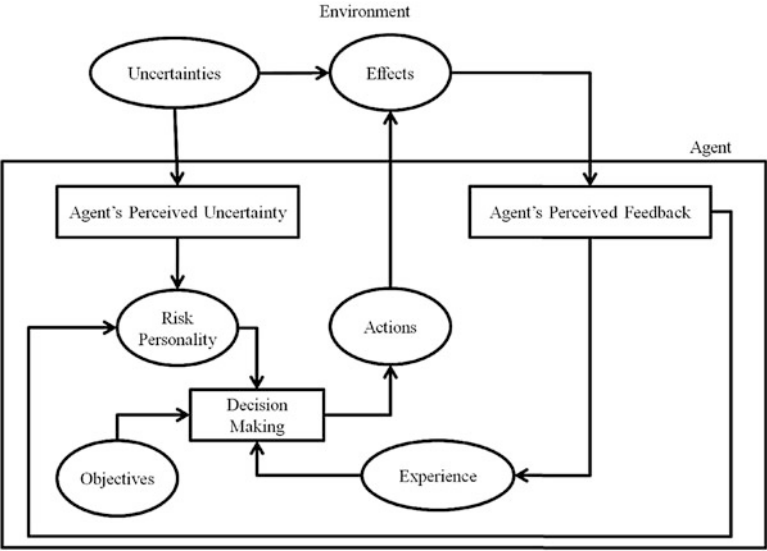


Fig. 2.6 The role of uncertainty in an agent’s decision making cycle

When the action is executed, an effect is generated in the environment, which the agent senses through its sensorial capabilities and feedback; this effect is then used for further learning. We note that this effect carries uncertainty information as well.

The cycle continues, and the agent continues to perceive the uncertainty in the environment, evaluating its impact on objectives, producing an action accordingly, monitoring the effect, and generating appropriate feedback to update its experience and learn.

The diagram shows that the agent’s risk was a function of its objectives and uncertainty.

Definition 2.9. Risk is the impact of uncertainty on objectives.⁶

The definition of risk above includes both positive and negative impact; therefore, it assumes that risk can be negative or positive. For example, the risk of investing in the stock market can be positive (profit) or negative (loss). In both cases, we would use the term risk because at the time the decision was made to invest, the decision maker should have evaluated both possibilities: the possibility of making profit and the possibility of making loss. An educated decision maker when making a decision to invest accepts the negative risk as a possible outcome, and equally, the positive risk as another possible outcome.

⁶We have changed the definition of risk from the one introduced in ISO3100 [8] by using the word “impact” instead of “effect”. The reason is that the word “effect” has a more subtle meaning in this chapter.

The common goal of a CRT exercise is to manage risk. This claim is safe because underlying every use of CRT discussed in Chap. 1 lies in objectives and uncertainties that derive the overall CRT exercise. The CRT exercise is established to fulfil a purpose that takes the form of a function. One of the main functions of CRT discussed in Chap. 1 is to discover vulnerabilities as a step towards designing a risk-management strategy. By discovering vulnerabilities, we become aware of them and we can take precautions to protect the system. However, what is a vulnerability?

ISO3100 defines vulnerabilities as “a weakness of an asset or group of assets that can be exploited by one or more threats”[8]. In this book, we will adopt a definition from a system perspective [4] because words such as “assets” can be confusing if they are not understood from an accounting perspective. As such, the following definition of “vulnerability” is provided.

Definition 2.10. A vulnerability is the possibility evaluated through the level of access or exposure a hazard or a threat has to a critical component of a system.

A hazard is an unintentional act that may harm the system such as a fire. A threat is an intentional act such as a hired hacker who has the intention to hack into the computer network and cause damage. For the network administrator, this hacker is a threat.

Vulnerability exists through exposure to an authorized or unauthorized (even accidental) access of a critical element to a hazard or a threat; we will refer to this exposure as “events.” What creates risk is the level of uncertainty of this exposure, and the magnitude of damage that can accompany the exposure if it occurs; thus, the uncertainty surrounding the circumstances in which the event will occur will impact the critical element, which will in turn impact the objectives.

$$\text{Risk} = \text{Vulnerability} \otimes \text{Effect}$$

The building blocks for hazards and threats are shown in Fig. 2.7. These building blocks provide knobs to control hazards and threats. An entity needs to be capable of performing the act. Therefore, capability is one building block. We will revisit the concept of capability and deconstruct it into components in Chap. 4. For the timebeing, an entity has the capability if it has the ingredients to provide it with the capacity to perform the act. For example, a computer hacker needs to have the knowledge to hack into a computer. In Sect. 2.14, we will call this know-how the skills to hack into a computer. The collective skills necessary to perform the act of computer hacking represent one dimension of the capability of the entity. Similarly, for a bushfire to ignite by nature, the ingredients of the capability need to be in place. These can be the ability of the environment to have high temperature, dry weather, etc. A thief who is denied the knowledge to hack a computer can’t become a computer hacker because the thief was denied the capability.

While we will expand more on the concept of a capability in Chap. 4, we will approximate the ingredients of a capability in this chapter to physical ingredients and know-how ingredients. Most of the analysis conducted in this book will focus on the know-how. This is on purpose for two reasons. First, without the know-how, the physical ingredients are insufficient. While it is true also that without the physical

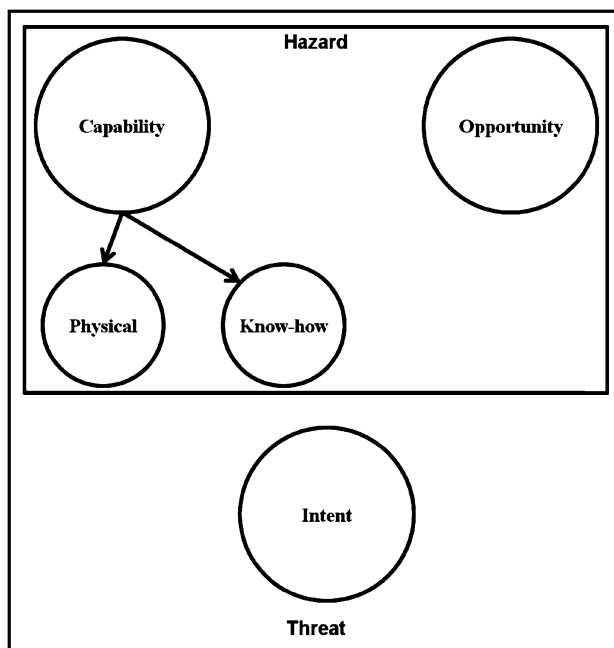


Fig. 2.7 Building blocks of hazards and threats

ingredients, the know-how is insufficient, but the know-how is more important because it can identify different ways of designing the physical ingredients. Second, since CRT is mostly about threats and threat-actors, the know-how shapes up the behavior of the threat actor, and the characteristics of the threat.

The opportunity is about the exposure component in the definition of a vulnerability. A computer hacker who is denied access to the computer network has been denied the opportunity to hack into the network, despite that the hacker has the capability to hack. The black box in an airplane is robust against high impact collision and fire so that in aircraft accident investigations, the recording can be replayed to shed light on the accident. By placing the recording device inside the black box, fire and collision as hazards or threats have been denied access, therefore, have been denied the opportunity, to cause damage to the recording.

Therefore, regardless of whether we are talking about hazards or threats, both the capability of the entity and the opportunity need to exist. Moreover, in the case of a threat, intent is needed. A computer hacker who has the capability to hack into a network, and has the opportunity by being left alone in the network room without any surveillance can hack into the network if the hacker wishes to. At this point, the intent of the hacker is the only thing between hacking the network and not hacking it.

The three building blocks: capabilities, opportunities and intents, are key in any risk assessment analysis because they offer tools to analyze complex systems, while also offering structured ways to think of the remedies. As the example above illustrated, to eliminate a threat, one can deny knowledge as a mean to deny

capability, one can deny access as a mean to prevent exposure and, therefore, the opportunity to create an impact on critical elements, and one can shape and reshape intent so that entities with the capabilities and opportunities do not become threats in the system. This type of analysis can be used to assess the risk accompanying the different roles of a red team that were discussed in Sect. 1.6.2.

Let us now take a more complex example that mixes hazards with threats. Assume a system user who leaves their password on their mobile telephone to remember it, the mobile telephone is stolen and a criminal uses the password to break into the system. In this case, the user did not have the intention to cause damage, despite this possibly being considered an act of negligence. While the password was the means to obtain unauthorized access to the system through the intentional act of the criminal (a threat), the availability of the password to the criminal was not intended by the user (a hazard).

A critical component such as the heart in a human becomes a vulnerability when it is exposed to a hazard such as a car accident or a threat such as someone intentionally attempting to dysfunction the heart through a stab wound. The vulnerability here arises from the level of access that was granted to the hazard or threat by the holder of the critical element. If a fence was built that was capable of stopping the car from crashing with the human, access has been denied, and therefore, this particular vulnerability has been eliminated.

Before this discussion ends, one final definition is necessary. This definition is often ignored in risk-management literature-the definition of a “trigger.” It must be understood that the event would normally require a trigger. A trigger is a different type of event. Becoming angry with someone may trigger violence. The event of violence would expose some critical elements of the system to a hazard or a threat; thus, creating a situation of risk.

Here, the word “trigger” is preferred over the word “cause.” A strict definition of a cause is that the effect would not materialize without the cause. If someone is angry, many things (i.e. triggers) can happen to make this person produce an undesirable action. More importantly, these things can happen still and the effect may not occur. None of these things is a cause per se; the real cause is the cause for the person’s anger, which could have been that the person failed an exam. Therefore, a trigger can be considered an auxiliary cause or an enabler for the effect to materialize [1].

For example, if throwing a stone at a window causes the glass to shatter, the effect of the action is shattering. Before the action is produced, the effect of the action must be evaluated while considering the possibility that the force of the stone is not sufficient to cause the window to shatter. Thus, uncertainties should be considered when evaluating expected effects.

We will avoid discussing causality in its philosophical form. Despite the fact that some of these philosophical views are the basis for some of the tools used in this book, they are not essential for understanding the materials in this book. Interested readers can refer to [1].

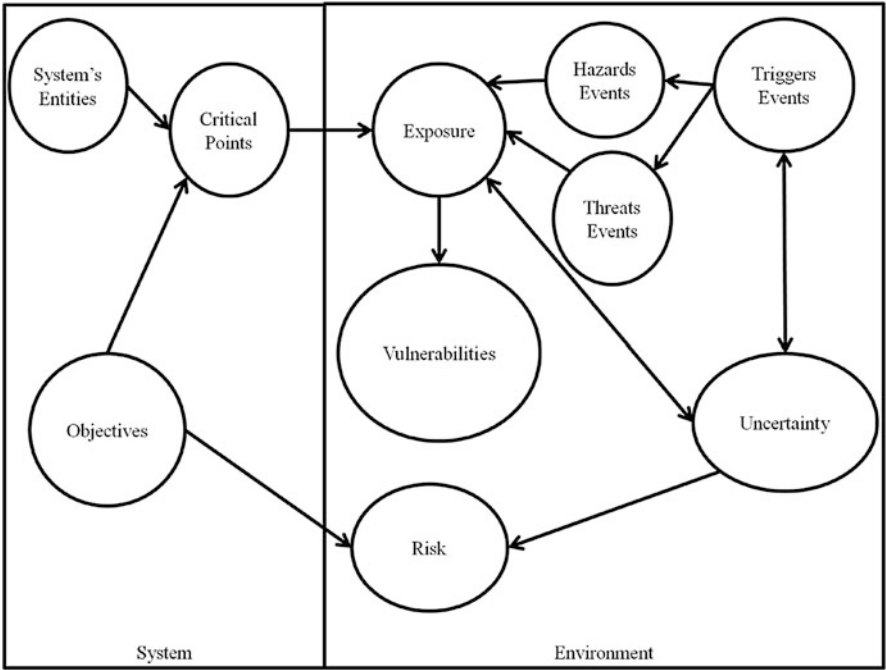


Fig. 2.8 A diagram connecting different concepts related to risk

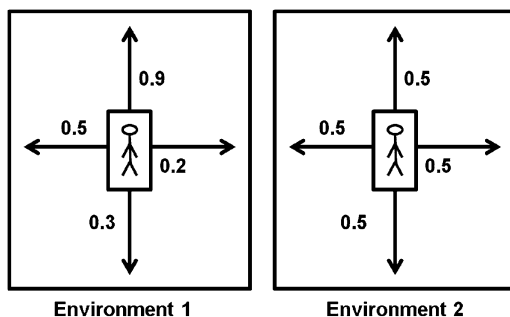
It is important to understand the difference between a trigger event and a hazard or threat event because trigger events are those events that we need to control to prevent a hazard or threat event from occurring. Figure 2.8 captures these concepts in a schematic diagram [11].

In Fig. 2.8, the rectangle on the right-hand side represents the elements of risk that lie in the environment, while the rectangle on the left-hand side represents the elements of risk that lie within the system. This distinction is critical when analyzing a system.

As discussed at the beginning of this chapter, blue sees red as a system and vice versa. The blue system sees red as one of the systems in the blue’s environment. It represents sources of uncertainty to blue. There can be internal sources of uncertainty within the blue system. However, blue would believe that internal sources of uncertainty such as questions about whether their own internal team is ready to face much larger jobs than what they currently perform are more controllable through something such as training than external uncertainties.

The fact that blue sees red as one of the systems in the environment is only half of the story. In CRT, blue needs to see itself as an external system to red in red’s environment. Consequently, as red shapes blue’s space of unknowns, blue also shapes red’s space of unknowns. These mutual interdependencies are the essence of the interaction between red and blue. They define a space of uncertainties, where

Fig. 2.9 Role of uncertainties and objectives in different environments



uncertainties are interdependent. Uncertainties in CRT are no longer external factors to the system, but tools and knobs that a system can control and use for its own advantages.

Blue should not act in a passive manner, accepting what red does as sources of uncertainty. It should take active steps in shaping its own actions so that it increases red's uncertainty. As uncertainties increase, red is likely to ignore its original objectives and make uncertainty management one of its objectives. Red will be overwhelmed with uncertainty, making the likelihood of an outcome a great deal smaller so that differences because of magnitude of outcomes become irrelevant.

To illustrate this, consider the case in which you wish to invest your money. You have a number of options with a reasonable level of uncertainty and a possibility of a high return. The decision you would make is simple: select the combination of options that maximizes your expected return. The expected return is a combination of uncertainties that impact each option, and the ability of each option to maximize your objective of achieving maximum return.

Now, consider the case in which all options with which you are faced have very high uncertainty. As uncertainty increases, differences between the options based on return decrease in the eyes of a human. Uncertainty takes over. In this case, you must not focus on return.

Instead, you should identify for which option you can control or influence uncertainty for your own benefit. You will focus on managing uncertainty, making controlling uncertainty your objective.

Figure 2.9 illustrate this counterintuitive point. Imagine an agent is attempting to find an apple in the two environments shown in the figure. In Environment 1, the agent has four roads to follow, and each road has a probability of finding an apple associated with it. The three directions of North, East and South have low uncertainty (high or low probability), while the west direction has high uncertainty where it is a 50–50 chance to encounter an apple. In this environment, the agent needs to focus on the goal of finding the apple. A rationale agent will start with the north direction as it offers the highest probability for encountering the apple.

In Environment 2, the situation is different. All four directions have high uncertainty. Classic decision making would suggest to start with any, since the expected value is equal for all four directions. In RT, however, we understand that uncertainty should not be a fact that we must obey. Instead, we can challenge

the uncertainty by seeking more information. In this situation, the agent changed the objective from maximizing the expected value for finding the apple to minimizing the uncertainty in the environment. When the agent manages to minimize uncertainty, the agent becomes ready to shift its focus back to maximizing return.

Controlling uncertainty is a non-intuitive concept. In almost all types of classical modeling presented in the literature, the emphasis is placed on how to represent uncertainty and incorporate it in the model so that the solution produced by the model is robust and resilient against the uncertainty. That is, classical modeling approaches uncertainty from a passive perspective, seeing uncertainty as external to the system, and the responsibility of a system's designer is to find designs and solutions that can survive the uncertainty.

CRT has a different perspective on the concept of uncertainty. Through CRT, we can see uncertainty as a tool. Red must realize that through its own actions, it can maximize blue's uncertainty. Blue needs to realize the same. Red can confuse blue and blue can confuse red. This form of a deliberately designed deceptive strategy is not about deceiving the opponent team so that it believes one thing will be done while the intention is to do another. Rather, deception here denotes deceiving the opponent to the point at which they do not believe anything. The opponent becomes overwhelmed with the uncertainty in the environment to the extent that it becomes paralyzed. It does not move because every possible direction in which it can move is full of unknowns. In such situations, the opponent will either not move at all or will simply make a random move.

A CRT exercise takes an active approach toward the discovery of vulnerabilities. In the majority of the CRT exercises, even if the individual exercise is concerned with the discovery of vulnerabilities caused by hazards, the issue of "intention", therefore "threats", demands a different type of analysis from that which involved with hazards. A criminal breaking into the system, after obtaining access to the password through the mobile telephone is an intentional act. This act becomes deliberate when it is planned. Studying the interaction between objectives and uncertainties is the key difference between what we will term an "intentional action" and a "deliberate action." This difference may appear controversial from a language perspective given the two concepts of intentional and deliberate are synonymous in English, and are used synonymously in many textbooks. However, here, we highlight differences between the two words.

2.2.5 Deliberate Actions

Within the class of intentional actions, we will pay particular attention to the subset of deliberate actions. We will distinguish "intentional" from "deliberate" to differentiate between classical decision making in an environment in which risks are not consciously evaluated by a red teamer (but in which the actions are consistent with the intention of the person) and decision making that is always accomplished after careful risk assessments.

Definition 2.11. A deliberate act is the production of an intentional act after careful assessment of risk.

In classical AI, the term “deliberate action” implies an action that has been decided on based on the construction of a plan. The definition we use above is more accurate because the emphasis is placed on risk assessment; therefore, a plan is being produced with risk as the focal point for evaluating different options and decision paths.

Therefore, every deliberate act an agent generates should contribute to the objectives. A series of effects is usually required for an agent to achieve one or more objectives. These objectives in their totality should reflect and be aligned with the purpose of the system.

In CRT, the impact of the uncertainty surrounding deliberate actions is evaluated on both red and blue objectives (i.e. self and others). Because the actions are deliberate, part of the CRT exercise is for each team to assess and analyze the actions of the other team. By analyzing actions, one team can reveal intent, drivers, objectives, and even the perception of the other team of the uncertainty surrounding them.

The previous statement should be read with a great deal of caution because of two problems. The first problem is that we can become so overwhelmed with analyzing actions that we utilize almost all resources without reaching any end. The second problem is that actions can be random and/or deceptive on purpose; therefore, a naive analysis of actions can mislead and counteract the CRT exercise.

Let us revisit the first problem. Some extreme views may perceive that there is an intent behind each action. This might even be misunderstood from our discussions above. We need to remember here that we are not discussing human actions in general; we are discussing actions within the context of the CRT environment. Therefore, there is a level of truth that we should expect that actions are produced to achieve intent. However, the true complexity here lies in the fact that to achieve one intent, there might be a need to design a number of actions. Some of these actions need to be generated in sequence, while others do not depend on any order. This defines a critical problem where the intent of the agent from a series of actions need to be inferred. This is a difficult problem requiring advanced techniques from the field of data mining. An introduction to data mining will be given in Chap. 3.

The second problem mentioned above is that actions can be deceptive and/or random. An agent may produce random actions to confuse the other agent. Here, the concept of deception is paramount and greatly impacts the behavioral data-mining methods. We may think this is becoming too complex. We may feel the need to ask how we can discover intent when deception is used. It can be surprising to learn that deception can actually help us to discover intent. If we consider the fact that deception in its own right is a set of deliberate actions designed to lead to an intent that is different from the original intent, we can see that the intent inferred from deception can give us an idea of where the real intent of the agent is. Of course we need to ask ourselves how we would know in the first place that these actions were

designed for deception and how we could categorize deceptive and non-deceptive actions. This is when complex tools, algorithms, and human's educated judgements blend together to answer this question.

2.3 Performance

Ultimately, in CRT the aim is to challenge the performance of a system. This task itself can take many forms, from a grand strategic vision on improving the economic performance of a country, to an efficient optimization algorithm of the bus system in a city, to a controller navigating a robot in an urban environment or a big-data mining algorithm to detect zero attacks in a computer network. Performance is a fundamental issue that we need to understand before discussing a theory of challenge. In effect, we challenge performance; therefore, we need to understand what performance means, what are the ingredients of performance, how to measure performance, how to analyze performance, and how to shape and reshape performance when we challenge performance.

To understand performance, we need to delineate the building blocks and concepts underpinning performance. A good starting point for this discussion is the meaning of the word "behavior".

2.3.1 Behavior

For an agent to produce effects, it needs to act. The set of actions generated by an agent define what we will term the agent's "behavior".

Definition 2.12. Behavior is the set of cognitive and physical, observable, and non-observable actions produced by an agent in a given environment.

We could define behavior simply as the set of actions produced by an agent. However, this definition lacks precision and essential details. It lacks precision because an agent does not act in vacuum; an agent acts within an environment. First, let us define the environment.

Definition 2.13. An environment for an agent A consists of all entities that reside outside A , their properties and actions.

Therefore, the environment represents the wider context within which an agent is embedded. An agent is situated within its environment. The agent receives stimuli from the environment, generates effects in response, and continues to monitor the impact of these effects on those environmental states to which the agent has access.

Behavior is not limited to the physical actions produced by an agent's set of actuators. Most of the physical actions are expected to be observable from an external entity. However, there is a group of actions that is generally unobservable;

these are the cognitive actions: the thinking process an agent experiences to reach a decision. Cognitive actions represent a critical component in an agent's behavior. We cannot simply ignore them because they are hidden in the agent's mind. In fact, if we can learn how an agent thinks, or at least the drivers behind an agent's decisions, we can predict most intentional physical actions. However, achieving this is extremely complex.

Meanwhile, one can see physical actions as the realization of cognitive actions. Walking to the restaurant to propose to my partner is a set of physical actions. These physical actions indicate that I have thought about the decision, and made a commitment to execute the action of proposing, with the expectation that the effect of marriage will become a reality.

The interplay between cognitive and physical actions is important in CRT. Once more, it is important to remind the reader that we are not discussing actions in life in general; this is all within the context of CRT, that is, an exercise with a purpose. Let us consider two examples at the two ends of the spectrum of CRT: one in which we are red teaming a strategic scenario on a country level and the other in which we are red teaming a computer algorithm for encryption.

In the first example, analyzing the cognitive actions of blue is about understanding factors such as how the blue team plans, evaluates options, and makes choices. These cognitive actions can be inferred, with different degrees of difficulty, from the physical actions of the blue team. For example, the division of the budget between buying capabilities to conduct cyber operations and buying tanks would provide us with an indication of how the blue team is thinking, where they see their future operations, and what possible strategies they have to meet their future uncertainties. These actions are not created for deception. It is less likely that blue will invest billions of dollars in tanks simply to deceive red; the scarcity of resources as a constraint reduces the space for this type of deceptive actions.

In the second example, the cognitive actions represent how the encryption algorithm thinks internally, that is, how it performs encryption. If the algorithm is an agent, we can notice its input and output. Breaking up the algorithm here is to uncover the computations it uses to transform this input to that output. We are attempting to use the external physical actions to infer the internal cognitive (problem solving) actions of the agent; by doing this, we can evaluate the robustness of our system, which is using this algorithm for storing data against attacks.

2.3.2 *Skills*

A red teamer attempts to interfere with, influence and shape the blue team behavior (action space). Therefore, for the blue team, the red team is part of blue's environment. Similarly, for the red team, the blue team is part of the red's environment. The red and blue environments share common elements: the shared environmental components between blue and red, and the components forming the interface between blue and red.

As red attempts to impact blue, it needs to rely on this interface, that is, the shared subset of the environment to generate effects. The ability of either team to act to generate effects on the other depends on their skill level.

Definition 2.14. A skill is the physical and/or cognitive know-how to produce actions to achieve an effect.

A skill is about the know-how related to achieving an effect. Some may define skills as the know-how to perform a task. However, here, the concept of a task is very limiting. By focusing on know-how for achieving an effect, we have a more flexible definition for “skill.” This definition links the outcomes (effects) to the processes and cognitive means (know-how). More importantly, by defining skills from the effects perspective, we emphasize that the agent’s choice of which know-how to use is based on the effects the agent wishes to generate, not on what the task that is being assigned to the agent intends to achieve. This is a crucial distinction for designing deliberate actions.

A skill cannot be defined in isolation; it always needs to be linked to a specific effect. However, effects have different levels of complexity and are generally nested. For example, the effect of producing on a computer a good essay based on recounting real events, while adding some details from the authors’ imagination, may be completed using different skills. Each of these skills link some level of know-how to an effect. One effect might be turning the computer into an “on” state (i.e. turning on the computer or ensuring that the computer is already turned on). This effect requires the know-how for sensing whether the computer is on. If the computer is not on, the know-how must be for sensing whether the computer is plugged in and that there is an electrical current reaching the machine as indicated with the power light, then using motor skills to press the “on” button. Another set of effects might be the production of a letter on a screen (this requires the know-how for generating motor actions to turn on the computer and press buttons); the effect of recounting the event (this requires the know-how for writing an account); and the effect of deviating from the actual story to an imaginary set of events (this requires the know-how to produce imaginary events in a coherent, interesting and engaging manner).

Each example of know-how listed above is composed of hierarchical knowledge divided into subsets of know-how. For example, the know-how to produce a letter on a screen requires the know-how of the layout of the keyboard; the know-how to translate the intent to write a letter (a cognitive event) to a series of muscle movements to press the buttons; the know-how to synchronize the fingers such that the correct finger is associated with the correct key on the keyboard.

The above level of deconstruction may seem as though it has too much detail. However, in CRT, the right level of detail will always depend on the objective of the exercise. If it is desirable to establish a writer profile to authenticate a person on a computer network, this level of detail will be appropriate.

In this situation, we need to know which fingers the person usually uses, and which keys are associated with which fingers. These two pieces of information (fingers used and finger-key association), together with the layout of the keyboard,

will provide us with an estimate of the time spent between pressing different buttons. For example, if a person uses only two fingers, one would expect a larger delay between pressing letters “h” and “o” when typing “hooray” as opposed to the delay between pressing letters “k” and “o” when typing “Hong Kong.” This information can establish a different profile for different users, which is then used as a background process for authentication and user identification.

Therefore, sometimes a level of detail for one exercise is not required for another. This is a decision that the CRT analysts must make.

A set of “know-how” forms a skill to achieve an effect. However, effects are hierarchical. Synthesizing effects on one level of the hierarchy requires specific skills (i.e. know-how to achieve a larger effect on an upper level of the hierarchy). It is important to recognize that it is not sufficient to take the union of the skills required to achieve the low-level effects to achieve the higher level effect. We need to ensure that we also have the know-how to synthesize the low-level effects. Therefore, the whole is not the sum of the parts.

This discussion indicates that skills are organized in a hierarchy, which is a commonly accepted notion in information processing and behavioral sciences. The challenge of a discussion such as this for CRT activities is that we can continue deconstructing a high-level planning task (as in the case of planning the cultural change required to accommodate next generation technologies in a society) into smaller and smaller tasks down to an arbitrarily microscopic level. The main question is whether this helps?.

In a CRT exercise, we need to deconstruct down to a level after which further deconstruction of skills is not needed. Therefore, the concept of a skill as defined above offers the red teamer a critical dimension for analysis. By analyzing the blue team’s skills, red can evaluate blue’s limitations, discover its vulnerabilities, and can reshape its own environment to generate innovative effects far away from the know-how of blue. Red can even help blue by designing training programs to improve blue’s skills in specific areas so that blue generates effects that are useful for them but are far away from those in which red is interested. As long as we avoid deconstructing effects and skills beyond what is appropriate and useful for the exercise, this type of deconstruction is vital for the success of the analysis.

2.3.3 Competency

An agent’s behavior is defined by the actions the agent produces; these actions are the product of the agent’s skills. There is a direct relationship between skills and behaviors. An agent uses its know-how to generate actions to achieve effects. The totality of these actions represents the agent’s behavior. Thus, an agent’s behavior is the product of the agent’s cognitive and physical skills. However, how can we evaluate behavior or skills?

Definition 2.15. Competency is the degree, relative to some standards, of the level of comfort and efficiency of an agent in adopting one or more skills to achieve an effect.

Competency is the measure of performance we will use to assess an agent's behavior. It acts as an indicator for the nature of the know-how (skills) an agent possesses.

The definition above requires further discussion related to two factors, the need for a standard to measure competency, and the distinction that has been made between comfort, which is a characteristic of the agent, and efficiency, which is a characteristic of the task.

2.3.3.1 Need for a Standard

Competency is measured relative to a standard, which is a reference system against which one can compare agents. Gilbert [5] uses the elite or the fittest in a population as the reference point. (Section 2.3.4 will discuss Gilbert's views further.) Here, the main point to emphasize is how to set a standard in a CRT exercise.

It must be remembered that the red team comes from the blue team's culture. As a result, the setting of a standard can be misleading if the standard is set relative to blue without knowing the standard for red. Let us consider two previous examples: one related to a strategic decision, while the other related to computer security. Assume two countries: X and Y . Country X is technologically savvy, developed, has a democratic government, and the population is highly educated. Country Y is undeveloped, relies on very old and outdated technologies, suffers from internal political instability, and the education level is poor.

It is clear in this example that if Y wishes to red team X 's plans, Y is incapable of evaluating X 's competencies. A red team drawn from Y for the purpose of thinking like X does not have the knowledge to do so, neither does it have the intuition and thinking abilities to imagine what are the right standards to use to evaluate X 's competency. Most likely, in this case, the exercise will fail because it is too imaginary or the exercise will simply be counterproductive.

However, does this mean that X can red team Y given that they possess the knowledge and technology? Let us assume X is attempting to conduct a CRT exercise to evaluate how Y will respond to an economic pressure that X will enforce on Y to steer Y toward becoming a democratic country.

In this example, we would need to ask what standards the red team should be using within X to assess the competency of Y in responding to this situation. It would not be surprising to expect that the standards used by X to evaluate Y are likely to overestimate what Y can do. The red team in X is highly educated. They can run scenarios on large computer clusters to evaluate every possible response that Y can produce. They have sufficient knowledge about Y that they can assume that they are able to think like Y . When assessing the competency of Y in applying

a specific policy tool to the situation, they can lower their own standards but realistically, having a very high standard is not a problem in this situation. So, what is the problem?

The main problem in this situation is that people in Y are extremely competent in a group of skills that X does not have. It is the know-how to use simplicity to respond to technology savvy know-how. Therefore, for X 's CRT exercise to be effective, X needs to accept that they may have a blind spot in their understanding of the behavioral space of Y . As such, how can red in X define the standard for this blind spot? There is no single answer to this question. The obvious answer is to study Y to the greatest extent possible. Providing the complex answers to this question is beyond the scope of this book.

Let us take a more objective example. Assume a group of thieves would like to rob a bank. The bank establishes a red team from their high-tech departments to identify the vulnerabilities that the thieves may exploit. The red team properly evaluates the competency of red in terms of every skill required to break into their computer network. The red team uses their standards to break into the computer network as the standard to evaluate the thieves' competency level. Let us assume that the thieves are not as skilled in cyber espionage, cyber strategies, computer security, and network intrusions. In such a case, the standards used by the red team remain appropriate, despite the fact that they are well above the level of capability of the thieves.

However, the thieves' objective is to rob the bank, not to break into the bank's IT network. Given that we are assuming that breaking into the IT network is a necessary condition for the thieves to rob the bank, it is fair for the red team to evaluate this as a vulnerability. However, the thieves do not have the know-how to break into the network. Instead, the thieves know how to blackmail, exert inappropriate pressures, and use violence and force. The thieves are not scared of breaking the law. Their behavior is embedded in an entirely different behavioral space from the highly educated IT team in the bank.

As such, the primary problem is that the skill space for the thieves cannot be fully discovered by the red team in the bank. Given that skills are nonlinear, there is no guarantee that the standards used by the bank are high enough to assess the competency of the thieves. The thieves may simply cause damage to the electrical power supply in the city, cause damage in the bank's computer system, force the bank to switch to manual operations, and steal the car with the money like in the old movies.

Setting a standard to define competency in CRT assumes that in a normal setting behavior is symmetric. CRT addresses symmetric and asymmetric situations; it is in asymmetric situations that setting standards relies on correctly mapping the behavior and skill spaces, and having the know-how (required skills) to set the standards properly.

How can we then establish standards in CRT? First, we need to change the standard from a ceil that defines a goal to a baseline that defines an objective. By using the concept of the elite, we establish an upper boundary on what can be achieved, we then attempt to measure how far the agents are from this upper

boundary (goal) based on the agents' outputs. However, the red team may not have the skills or knowledge to estimate this upper boundary properly. Overestimating the upper bound is not necessarily a bad thing, but arbitrary overestimating this upper boundary in an ad-hoc, blind manner or underestimating it are real vulnerabilities for the CRT exercise because the other team might be greatly more competent than what the red team think.

Moreover, a ceil is established under the assumption that we know what the effect space is. In the absence of complete knowledge of the effect space, we cannot define this ceil. Therefore, in CRT, we need instead to move away from this idea of establishing the standard as a ceil. Instead, competency of one team will be defined relative to an assessment of the performance of the other team. We term this "comparative competency."

Definition 2.16. Comparative competency is the degree of the level of comfort and efficiency of an agent in adopting one or more skills to achieve an effect in one team relative to the ability of the other team in achieving the same effect.

In comparative competency, a team expresses its competency relative to the performance of the other team. Therefore, competencies are expressed as two percentages, one related to the comfort of red relative to blue, and the other related to the efficiency of red relative to blue when attempting to achieve a particular effect.

Comparative competency does not address the problem that one team may have a blind spot in mapping the other team's skill space. This problem requires multiple treatments, especially with regards to team membership discussed in Sect. 1.3.

Remember that different skills can come together in different ways to achieve the same effect. Therefore, when measuring competency, we are measuring to the best possible performance that the other can display in achieving the effect. Since this best possible performance is dynamic within a CRT context, because of the learning occurring within the CRT exercise, comparative competency is a dynamic concept.

2.3.3.2 Comfort vs Efficiency

Given that we will explicitly distinguish between the cognitive and physical attributes of, and functions performed by, agents, it is also important to distinguish between comfort, the level of ease in achieving an effect, and efficiency, the accuracy and speed in achieving that effect.

Imagine you are at the checkout counter of a supermarket. The cashier behind the counter is scanning your items, and placing them in a bag. One of the cashiers might be the elite in that supermarket because every item they place in a bag is scanned (100 % accuracy) and they can scan and package 20 items per minute. This cashier is defining the standard for the checkout counters in this supermarket.

Judging on throughput alone is not sufficient for us to understand the long-term effect. The level of comfort, the cashier's feelings and perceptions about the ease with which they perform their job can provide us with a more informative picture of performance, and the ability to predict long-term effects. If the cashier perceives

that the job is very easy and simple, we may assume that their performance would degrade if they worked without rest for 1 h. If they perceive that the job requires a great deal of effort and they need to concentrate to ensure the accuracy of scanning and packing the items, we know that the cognitive load becomes an important factor in this situation and the cashier's performance may degrade in 30 min without a break instead.

This discussion emphasizes that competency cannot rely on agents' physical and observable actions alone, it should also consider the agents' cognitive actions. Whether or not to assess these cognitive actions requires cost-benefit analysis. A study needs to decide on the importance of this type of data to the particular exercise. Cognitive data can be drawn from questions posed to the subjects or from sophisticated data-collection mechanisms such as brain imaging. This is an exercise-specific decision.

2.3.3.3 Revisiting Behavior

We can now redefine behavior, or offer a second definition of behavior.

Definition 2.17. A behavior is the expression of an agent's competency level and acquired skills in actions.

In this definition, we emphasize competency (comfort and efficiency) and skills (know-how) when observing an agent's set of actions to discuss the agent's behavior. This definition, illustrated in Fig. 2.10, moves us a great deal closer to a useful definition of behavior that provides us with the tools for analysis. Competency provides the measures and indicators, while skills guide the data-collection exercise to focus on the how. By understanding the how, we can diagnose the behavior, and through competency, we can observe the impact of different treatments on behavior.

Figure 2.10 connects the concepts discussed so far in a coherent form, and introduce additional points for discussion. It differentiates between two types of knowledge. Axiomatic knowledge need to be acquired by an agent through transfer from another agent. We will reserve the function of education to the transfer of axiomatic knowledge.

Learned knowledge is acquired through mechanisms such as training, practising and challenges. Training assists the agent to improve efficiency on a task. Practising provides the agent with the confidence and comfort in performing a task. A challenge through any form of interaction, including training and practice, provide another mean to extend the agent's learned knowledge. These learned knowledge become an input to the agent's know-how knowledge base.

The agents' skills and competency come together to form the agent's behavior, which is expressed in the form of actions. Through self-reflection on these actions, as well as training, practising and challenges, the agent learns new knowledge.

In CRT, we will assume that everything is measurable, whether directly or indirectly, through a set of indicators that approximate the phenomenon we wish to measure. We also assume that everything is for a reason; therefore, there is a cause

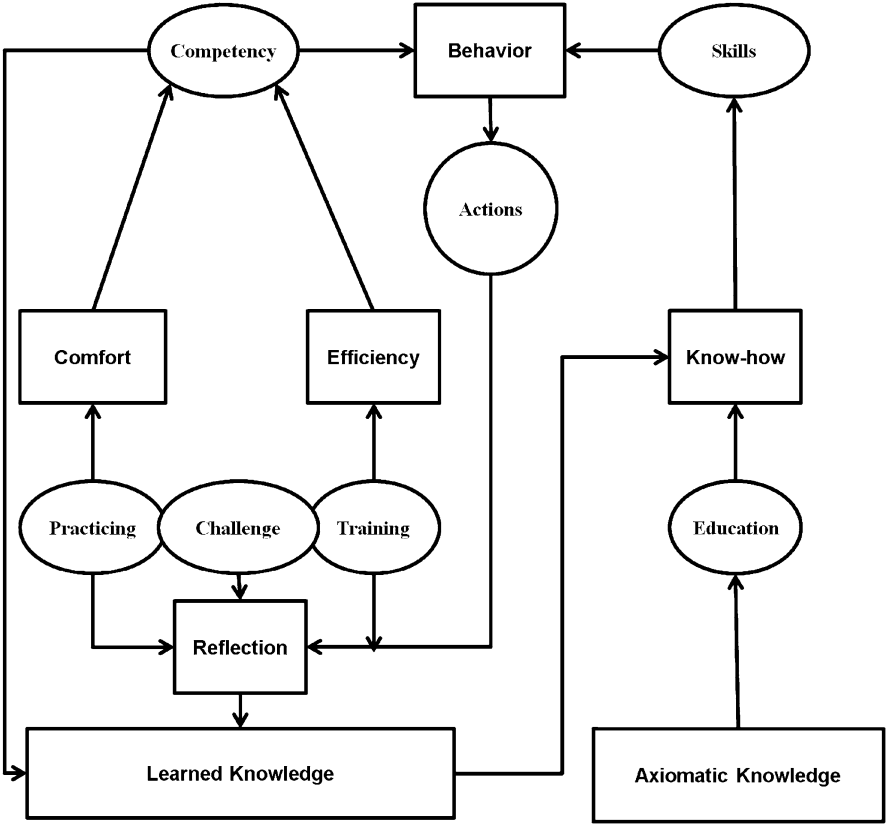


Fig. 2.10 Deconstruction of behavior

for everything (be it physical, cognitive, social, or environmental) underlying every piece of information in a CRT exercise, even if the purpose is to deceive the other team, or if the information was mistaken demonstrating a level of inefficiency in the other team. Without these assumptions, we cannot perform computations; thus, we cannot systematically analyze a situation. CRT evolves through measurements.

In CRT, we need to measure to compute, and we need to compute to measure. We need to understand causes to control and influence, we need to influence to create effects, and we need to create effects to achieve the purpose of the exercise.

The question we need to ask is how to measure behavior. If we wish to challenge behavior, we need first to be able to measure it. Otherwise, we will have no means

by which to establish with confidence whether we were able to challenge anything. Understanding behavior through skills and competency gives us a powerful tool to measure behavior; in fact, it is so powerful that through these factors, we can automate the processes of production of behavior and the analysis of behavior.

Any action producible by an agent in a CRT exercise does not occur in a vacuum. Both the agent's skills and competency level of these skills shape, influence, and even determine the production of an action. If an agent lies, the agent needs to have the skills of lying. If the agent lies well, the agent needs to lie well relative to the other team, ensuring that the other team believe the lie. If the agent falls because their feet are weak, there is a physical constraint limiting the agent from reaching maximum competency of the skill of walking. If the agent mumbles or produces a grammatical error, it might be caused by the fact that the agent's cognitive resources have been depleted; thus, the agent is making mistakes, resulting in lower competency with certain skills.

Thus far, we have discussed skills, competency and behavior. We have discussed competency as a means by which to measure performance. It is now time to discuss performance.

2.3.4 From Gilbert's Model of Performance to a General Theory of Performance

The model we will use in this chapter is inspired by the work of Gilbert [5], the father of performance engineering or what he termed "teleonomics." However, we will deviate from Gilbert's views in part to design views appropriate for the CRT context of this book, and to ground his system-thinking views in computational models.

Gilbert sees the world split into two components: the person (P) and the environment (E). We should recall that a person in this book can be a group or an organization. When the person receives a stimulus, they need to be able to recognize it. This recognition is fundamentally conditional on their ability to recognize the stimulus. Gilbert termed this "discriminative stimuli:" S^D .

When a person receives a discriminative stimulus, they need to have the capacity to respond. Gilbert termed this "response capacity:" R . A person may have the recognition system to receive and comprehend the stimulus, and the capacity to respond, but they choose not to respond simply because they do not have the motivation to do so. Therefore, the response needs to be accompanied with "stimuli reinforcement:" S_r , which for the person represents the feedback to their motives.

The above can be summarized in Gilbert's notations as

$$S^D \rightarrow R.S_r$$

Table 2.1 An example for mapping Gilbert’s model to a scientist job

	S^D	R	S_r
	Information	Instrumentation	Motivation
Environment	<i>Data</i>	<i>Instrument</i>	<i>Incentives</i>
	Literature	Functional Laboratories	Funding
Behavioral repertory	<i>Knowledge</i>	<i>Response capacity</i>	<i>Motives</i>
	Education and training (know to recognize)	Thinking and skills (know how)	Ambition

The → represents implication in his notational system. Gilbert then divided the environment into three components that correspond to the three components associated with a person: *data* represent the information delivered to the person through the stimuli; *instruments* represent the instrumentation component to deliver the response; and *incentives* represent the reward system to reinforce and/or trigger motivation.

We will present our own example below to explain Gilbert’s model, and to use it as the basis to explain other concepts in the remainder of this chapter. Let us take a scientist as the person we wish to model. Following Gilbert’s model, we can construct the matrix presented in Table 2.1.

The simple example presented in Table 2.1 demonstrates the strength of Gilbert’s model for CRT. First, the environment provides to the scientist the three enablers that allow the scientist to perform their job. The literature represents access to knowledge. For a scientist to innovate, access to the most recent knowledge that has been developed in the field is essential. If the environment denies the scientist such knowledge—for red teamers, this is a knob to achieve an effect of stopping the scientist from achieving their goal—the scientist might end up reinventing the wheel.

The instrumentation that the environment needs to make available to the scientist is represented here as the scientific laboratory and encompasses all the tools and instruments required for the scientist to do their job. Once more, if these facilities are not available, the scientist cannot produce the desired outcome and cannot materialize their ideas.

Incentive is a tricky concept in science. We would expect that a scientist requires some sort of incentive to perform their work. Here, we assume that incentives take the form of scientific funding and grants. These grants do not necessarily provide monetary incentive to the scientist, but a moral incentive, reflecting a recognition of the importance of the work. The monetary value can be used to improve the facilities and instrumentation; thus, speeding up scientific discovery.

The behavioral repertoire for the person captures the model of the stimulus-response discussed above. Here, we assume that the level of education and training provide the scientist with the ability to recognize stimuli in the environment. The author often says to his students,

You see what you know.

A scientist who does not understand mathematics will not be able to interpret an equation written on a whiteboard; thus, they cannot interpret the stimuli that may trigger an idea in their mind. Thus, education and training represent the knowledge repertoire required for S^D to function.

The capacity to respond for a scientist represents their thinking abilities and skills. To create new contributions, the scientist needs to have the skills and creativity to produce scientific outcomes from the stimuli. Their motivations are assumed to be internal and to take the shape of scientific ambition.

The model above gives us the basis to analyze the person from a CRT perspective, providing us with the knobs to influence performance and reshape it if needed.

The details of Gilbert's work can be found in his seminal book [5]; a very worthwhile read. His work is inspiring and well engineered. However, we need to search deeper and be more concise to transform his system into a system suitable for CRT. This is for several reasons.

First, Gilbert focused on a holistic view of performance, resulting in an efficient, but high-level, model that can guide human managers to improve performance. The objective in CRT is to challenge performance; therefore, we need to transform this holistic view into a grounded control model that enables us to steer performance to either positive or negative sides. Moreover, we need this model to be sufficiently grounded so that we can use it to compute, but not too grounded to avoid unnecessary computational cost.

Second, Gilbert did not seem to differentiate between the physical, cognitive and cyber spaces. By focusing on performance alone, it did not matter in his work whether the capacity of the agent was cognitive or physical, or whether the instruments used by the environment were psychological or physical. These elements are not included for the performance engineer to analyze based on the context in which they are working with. However, here, we prefer to make these distinctions clear given the tools and models to be used for CRT will be different.

In the example of a scientist, Gilbert's model is possibly useful for us as humans to see how we can manipulate performance from the outset. However, if red teamers wish to challenge this scientist with ideas, or challenge their environment to steer their scientific discovery one way or another, it is necessary to dig deeper. We need to separate the physical (e.g. laboratory) from the cognitive (e.g. creative thinking) and the cyber (e.g. access to information). Gilbert does this to some extent as we see in the example in which data and knowledge represent the stimuli, instrumentation represents to some extent the physical side, and motivation represents the cognitive. However, we can see also in the scientist example that this is not sufficient. A laboratory would have people such as post-doctorates and Ph.D. students who provide ideas to the scientist. These ideas can act as stimuli, responses or even motivations.

Third, Gilbert spent considerable time in his book as an anti-behaviorist. In fact, when one attempts to understand Gilbert's views properly, it is clear that he was not an anti-behaviorist because his own model, when analyzed properly, is a behaviorist model. However, it seems that the behaviorist laboratory in which he was raised, and the environment in which he was living were taking extreme views of the behaviorist approach. This caused Gilbert to take a strong stand in his book against behaviorism, while clearly his model demonstrated that behaviorism is embedded in the roots of his mind.

In our model, we will combine the cognitive, social and behavioral schools to provide multiple interception points a red teamer can use to understand, and if required, influence and reshape, a system. This will create a larger model, but one should zoom in and zoom out as needed based on the context, and the available data and resources. For example, Chap. 4 will present holistic models suitable for a strategist to use when designing strategies for a whole of government or for designing a strategy for policing and security. This will be in contrast to the type of data and the model used in Sect. 5.3 where we reach a level of detail on the level of human-brain signals.

Zooming in and zooming out in the model presented in this chapter provide the level of flexibility that a red teamer should have when analyzing tactical, operational, strategic, or grand strategic levels. Data, details, and even the models to be used are different, but the guiding principles and the thinking model are the same.

Figure 2.11 presents a complete rework of Gilbert's model, grounding it in the cognitive-science literature, or more specifically, information-processing research, while maintaining the underlying features to measure performance. We will first explain the nomenclatures below:

- U^c and U^p are the agent's cognitive and physical resources, respectively;
- L^c and L^p are the agent's cognitive and physical skills, respectively;
- E^c and E^p are the environment's cognitive and physical resources, respectively;
- A^r and E^r are the internal (agent's self-) and external (environment) rewards to an agent, respectively;
- M and I are the motives and intent of an agent, respectively;
- B represents the ability of the agent to perceive and recognize the stimuli; it is a fusion function of the stimuli, the agents' cognitive and physical resources, and the agent's cognitive skills;
- S and R are the stimuli in the environment, and the action/response produced by the agent, respectively;
- f is a fusion function that integrates the agent's response overtime and transform it to changes in the physical and cognitive resources in the environment, and into an environmental reward to the agent;
- O is defined as the positive or negative opportunity offered by the environment to the agent;
- $a \rightarrow$ in this diagram represents a flow of data, where the word "data" denotes elements such as information, knowledge, experience, and wisdom.

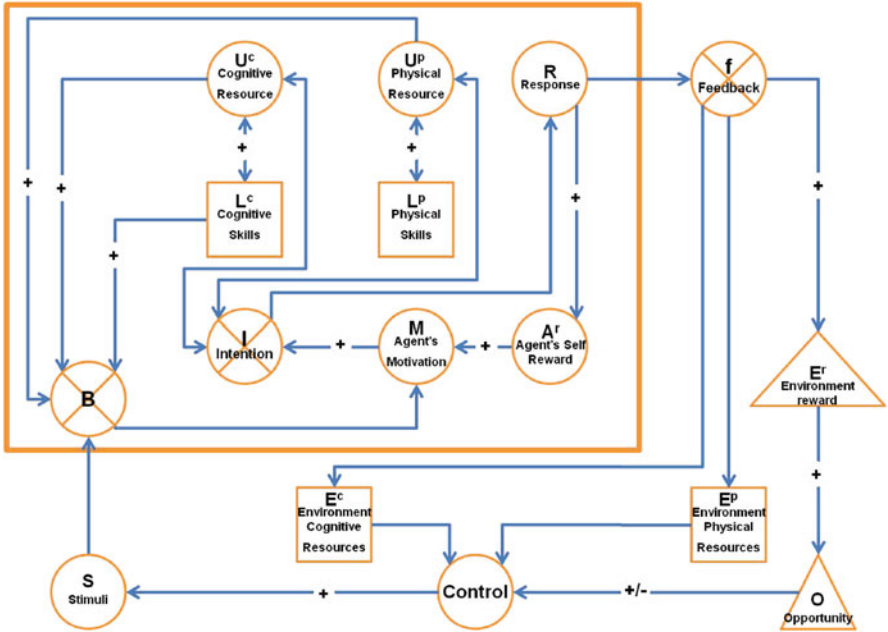


Fig. 2.11 An architecture of agent-environment interaction model

- + alone on an arrow represents a positive influence of change in flow and should be understood as the element at the tail of the arrow positively influences the flow to the element at the head. That is, if x_1 positively influences the flow to x_2 , when the flow to x_1 increases, the flow to x_2 is likely to increase and vice versa;
- +/- on an arrow represents a possible positive or negative influence of change in flow, that is, when the element at the tail of the arrow detects a change in flow, it may decide to increase or decrease the flow to the element at the end of the arrow;
- a circle shape represents a memory/storage + processing unit;
- a rectangular shape represents a storage unit alone;
- a triangular shape represents a decision unit;
- a shape of \otimes represents a fusion process, where one or more inputs need to be mixed and integrated over time to generate an output.

Let us consider the model to understand what it does and how it does this. The starting point is the circle labeled “control.” This circle represents the decision making that occurs within the environment, outside the agent. The environment may decide to generate a stimulus to the agent, either unintentionally (e.g. weather events) or intentionally (e.g. through the actions of other agents). A stimulus takes the form of a cognitive (e.g. data, ideas, experiences) or physical (e.g. information about a money transfer, or the information associated with giving a child a candy) flow. In this model, a flow is always a flow of data/information, although it may have a cognitive or physical source.

Once the information leaves the control point, it becomes a stimulus, S , to the agent. The agent receives this stimulus in a different form, B , from what it is in the environment. This form represents the fusion of different factors: the stimulus that was generated, the agent's physical resources, the agent's cognitive resources, and the agent's cognitive skills.

For example, if the agent lost the ability to taste (i.e. had a malfunctioning tongue), this limitation in an agent's physical resources would impact the agent's perception of tasting information in a stimulus. Similarly, if the agent is autistic, the lack of certain cognitive resources would impact the agent's perception of a hug. Finally, the agent's cognitive skills (e.g. the agent's knowledge of how to hug to reflect compassion or affection) would impact the agent's perception of a hug.

The perceived stimulus is then transformed into motives or goals. Sometimes, the stimulus may generate a new goal, as in the case of a new task being assigned to the agent and the agent needing to add to their repertoire of motives a new goal on the need to complete this task. At other times, the stimulus provides the agent with an update to one of its existing goals, as in the case of briefs from a subordinate that update the decision maker's knowledge of the rate at which existing performance indicators are being met.

The states, and the corresponding changes, of an agent's goals produce intentions to act. Intentions in this model are a product, not a system-state. The intention unit fuses the motives, the agent's cognitive resources, and the agent's physical resources to produce a plan. Information during this fusion process moves back and forth, where the cognitive and physical resources call and modify the cognitive and physical skills, respectively. During this process, cognitive and physical skills are updated and checked to produce the plan.

For example, assume an agent who used to be a professional swimmer had an accident in which they lost their right arm. Assume that the goal of the agent remains to be able to swim fast. Both the agent's cognitive and physical skills need to be updated. The agent needs to form a plan to move from the previous skills to a new set of skills that consider and match the new physical constraint.

The agent's internal plan can take the form of a series of actions that the agent needs to produce. However, only a limited number of responses can be produced at any point of time. Therefore, the intention unit also produces a schedule for the generation of responses. The first group of mutually compatible responses (e.g. a smile on one's face, together with a handshake) form a "response:" R .

The agent's internal response may be produced differently in the environment. For example, as the agent is moving their arm to shake a person hands tightly, the intended pressure on the other person hand is not properly produced. Thus, the handshake does not produce the intended effect.

Two rewards systems operate as action-production works. The first is the internal feedback, self-reward or self-punishment system in which the agent internally rewards itself. A person may attempt to reinforce their own goals to the extent that the person perceives that their goals are satisfied when they are not. This internal reward mechanism is very important because it is generally hidden and inaccessible from the outside world. It can act as a negative feedback cycle that

balances an individual's motives or a positive feedback cycle that cascades an individual's motives. When the internal reward mechanism gets damaged, it can create a personality disorder.

The second reward function originates from the environment, where other agents in the environment need to decide on the form and impact of such an environmental reward. We refer to this as the "opportunity:" *O*. The environment may offer or deny opportunities to the agent.

In our example of the scientist, the environment may decide to open the flow of physical resources-in this situation, the agent receives more funding; close the flow of physical resources-here, the funding stops; open the flow of cognitive resources-here, the ideas and knowledge produced elsewhere are communicated to the scientist (agent) to increase their knowledge repertoire; or close the flow of cognitive resources-here the scientist (agent) is denied such knowledge or cannot find people with the appropriate cognitive skills to extend their cognitive capacity to process information.

A red teamer's task is to understand how to design this control function, or at least influence it, so that the agent's actions and goals are consistent with the red teamer's goals.

A red teamer would aim to achieve one of the following three generic categories of goals:

- alter an agent's performance in a negative or positive direction;
- alter the effect of an agent's actions; or
- both of the above.

Figure 2.12 shows an abstract categorization of the role of CRT. In this categorization, CRT has two roles. One is to shape the actions of competitors, thus, the effectiveness of CRT is evaluated by measuring the distance between the intended effect of the competitor and the actual effect.

The second role of CRT is after the competitor's action, where CRT attempts to influence the effect generated by the opponent after it has been produced. Here, the effectiveness of CRT is evaluated by measuring the distance between the aspired effect of CRT and the net effect of the competitor. The competitor's net effect is the effect originally generated by the competitor minus the effect of the interference generated by CRT.

An example of the former role is when a red teamer wishes to alter the performance of a student to improve their performance in solving arithmetic problems. An example of the latter role is when CRT attempts to reshape the environment such that the advertisements a tobacco company is using have no effect on people.

When a red teamer aims at discovering vulnerabilities in a person, the primary reason that these are considered vulnerabilities is that the exposure of some critical elements to a threat will either impact the performance of the person or will simply impact the outcomes, and therefore, the objectives of the person. Each function of CRT discussed in Chap. 1 can be considered to achieve one of the three goals noted above.

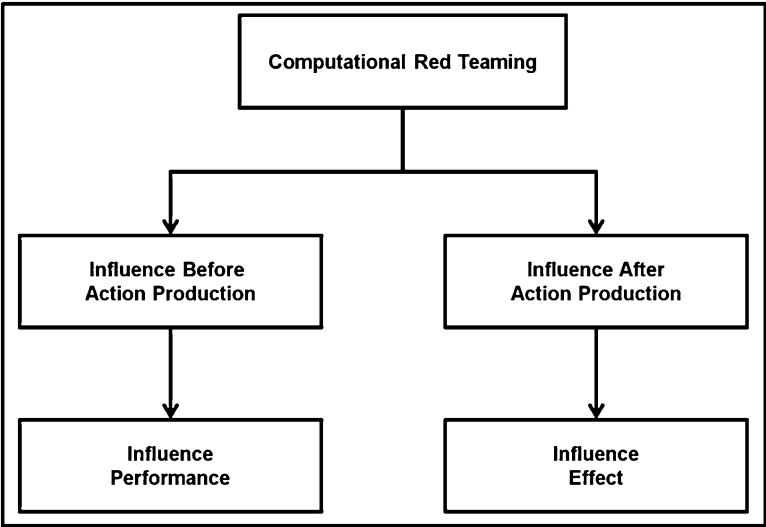


Fig. 2.12 Abstract categorization of the role of computational red teaming

The primary design mechanism a red teamer uses to alter performance or change the effects of blue is by designing a challenge. This is discussed in details in the following section.

2.4 Challenge Analytics

2.4.1 *A Challenge is Not a Challenge*

It is very difficult to isolate the very few scientific articles discussing the concept of a “challenge,” from the countless scientific articles using the concept to indicate a “difficult” or “impossible” situation. Therefore, it is logical to devote time here to explaining what a challenge is within the context of CRT. We need to go beyond a dictionary-level, common explanation of a challenge, to a more formal definition to ensure the possibility of designing models of challenge. We should see that a “challenge” here is a state that constructively achieves an objective. It does not denote the impossible, or a difficult situation.

Initially, we may see a challenge in simple terms: a challenge exposes an entity to a difficult situation. However, the main question is at which level of difficulty we are to employ the term “challenge.” It is very simple to ask other people difficult questions and criticize them for not knowing the answer, making them to feel inferior or incapable.

Take for example a situation in where parents would tell a 6-year-old child that they cannot earn money, that they are the ones who can buy the child what they want, and therefore, the child should listen to them. The child is exposed to what we would term in common English a “challenge.” They feel inferior in their ability to attract and own money. The child would be wondering what their alternative is to listening to their parents. The answer is obvious in this context; the child needs a manner in which to obtain their own money. The parents, without intention, generated an undesirable causal relationship in the child’s mind, that is, if the child was able to obtain money, the child could buy whatever they wanted, and therefore, the child could have an excuse for not listening to their parents. As presented below:

Obtain Money → Freedom to Buy Anything Desired

Obtain Money → No Need to Listen to Parents

These types of challenges are like a lose canon, they can fire randomly and even hit their owners. Within the scope of this book, we will not consider this example to constitute a challenge; we will simply consider it as an unthoughtful exposition to a state of hardship.

It is unthoughtful because the parents above would like to gain a position of power over the child as rapidly as possible. As a result, they state that if the child is unable to achieve something that they know or believe is far beyond the existing capacity of the child, the child must comply with certain conditions imposed by the parents. The parents fail to understand that this behavior may trigger a reaction of hostility and impose a feeling of hardship for the child. The child may rapidly adopt a hostile attitude toward their parents, or use their level of knowledge to find the quickest way to find money, which is obviously from the parents’ own pockets!

This is not the type of challenge we will model and discuss in this book. Instead, we will examine engineered, thoughtful and constructive forms of challenges whereby, the challenge is designed to achieve a desired outcome or effect.

2.4.2 *Motivation and Stimulation*

In computational sciences, the idea of how to create a stimulating environment has been studied a great deal in the areas of computational creativity, computer games, and simulation. Most studies in computational sciences on these issues are still in their infancy, sometimes offering conceptual models and ideas that are difficult to ground in a working system. However, and more importantly, “stimulating” should not be equated to “challenging” in a strict scientific sense.

Stimulating → *Challenge*

Challenge → *Stimulating*

The above notations emphasize that a stimulating situation does not necessarily mean that the situation was stimulating because there was a challenge associated with it. Similarly, a challenging situation does not necessarily stimulate the agent. An agent may be exposed to a properly designed challenge, but the agent may lack motivation or interest, which makes the situation less stimulating to them.

Criteria such as “stimulating” and “motivating” are more suitable for a human agent, as they require human traits and judgment. To generalize the concept of a challenge to a machine agent, we need to reduce these criteria to a set of indicators that can be used to assess and/or judge the process objectively without the need to rely on subjective judgment.

We use a simple to understand, but more complex to implement, criterion.

Definition 2.18. A task is challenging when the distance between the aggregate skills required to do the task and the aggregate skills that agent possesses is positive and small.

That is: $\text{Aggregate required skills} - \text{Aggregate possessed skills} > \varepsilon \rightarrow \text{a challenge}$ iff ε is small and > 0 .

We need to offer two words of caution here:

1. The concept of “distance” in the discussion above is not a simple quantitative metric.
2. The aggregate of skills is not the sum of skills.

Several sets of skills can be united in different ways to create different high-order skills. For example, let us assume that Jack is a creative person with excellent writing skills, and cryptographic skills. The skills of creativity and writing when put together may make Jack a creative writer. The skills of creativity and cryptography when put together may make Jack a good computer hacker. Practice plays the role of increasing the competency level of the agent. As the agent becomes competent, new skills emerge. As Jack practices his creative writing and computer hacking, he may develop skills in script writing for science fiction movies on quantum computations.

A good computer hacker is not created through simply by adding creativity and cryptographic skills. If it does, then we simply obtain two different people, one who is creative but has no understanding of computers, and the other who is a well-educated cryptographer but is not creative. When we put these two people together, it is unlikely that a good computer hacker will emerge for a long time, that is, the time required for each person to transfer some of their core skills to the other.

The above raises the important question of how to create a good computer-hacking team. The creative thinker needs to have some understanding of cryptography and the cryptographer should have a degree of creative-thinking ability or should be “open-minded.” There must be overlap of skills to establish a common ground for the members of the team to speak to each other in a language they can both understand, while not necessarily being an expert in the other’s field.

To recap the above discussion from a mathematical perspective, a distance metric on a skill space is not a trivial task, and the aggregation of skills is usually a nonlinear coupled dynamic system.

2.4.3 *Towards Simple Understanding of a Challenge*

There is not a great deal of literature on the concept of a challenge but there is small amount in the fields of education and psychology. Here, we will build on the work that does exist, but we must first deviate. As we will see, most of the literature treats the concept of a challenge in a holistic manner. A challenge is defined, then the concept is left to a designer such as an educator to interpret what it means within their context.

The online free dictionary [7] defines a challenge in many different manners. One definition that is particularly relevant to this book is the following: “A test of one’s abilities or resources in a demanding but stimulating undertaking.” This definition highlights the delicate balance that needs to exist between the two words “demanding” and “stimulating.” The need to strike this balance is supported by theories in educational psychology. Sanford’s theory of challenge is key in this area [10]. In his work, he explains the subtle difference between a challenge and a stress. He emphasizes the need to strike the right balance so that a challenge to a student does not turn into a stressful situation. This work was followed in the education domain by some scientific research on the topic [3, 9]. The pattern of a challenge was recently mentioned in a study on immersion, although there was no analysis of the specific pattern [6].

The above definition linked the concept of a challenge with the concept of “stimulating”. However, we separated these two concepts in the previous section. The word “demanding” is interpreted in our definition as exceeding the boundary. The word “stimulating” is interpreted that it is not too demanding to the extent that the agent may give up. However, the concept of “stimulating” has a second dimension related to the agent’s motives. A challenge would become stimulating if it has elements that triggers the agent’s motives. This dimension is agent-specific. As we discussed before, we separate the two concepts of a “challenge” and this dimension of the concept of “stimulating” in this book.

The concept of a challenge is traditionally found in the literature on “dialectics.” Naturally, considering a challenge can take the form of questions. Can we design a counter-plan for the opponents plan? Can we design an example to teach the students an extra skill they currently do not possess? Can we design an anomalous dataset that is sufficiently similar to normal behavior to be able to penetrate the anomaly-detection system for our testing purposes? Therefore, questioning is a natural mean to communicate a challenge. However, we should be cautious since not every type of questioning is a challenge. Questioning can be a mean for examination, interrogation and extraction of truth, sarcasm, or even a dry joke.

Mendoza [9] thinks of a challenge as “forcing myself to learn always to think at the limits.” Admiring the work of Ellul on dialectics, Mendoza cites Ellul’s four notions of a theory of dialectics:

1. Contradiction and flux are two characteristics in life that must be reflected in the way we theorize. Through the holistic approach of dialectic, meaning can be grasped.

2. The coexistence of a thesis and antithesis should not lead to confusion or one suppressing the other. The synthesis should not also be a simple addition of the two; instead, it emerges through “transformative moments” with “explosions and acts of destruction.”
3. The negative prong of the dialectic challenges the spectrum between the positive and negative prongs, creating change; or what Ellul called “the positivity of negativity.” Ellul sees change as a driver for exploration. Mendoza offers examples of the positives, including: “an uncontested society, a force without counterforce, a [person] without dialogue, an unchallenged teacher, a church with no heretics, a single party with no rivals will be shut up in the indefinite repetition of its own image.” [9]. These positives will create a society that resists change.
4. Automaticity of the operation of dialectic is not possible because many of the contradictory elements in the society are necessarily going to create those unique dialectic moments. Ellul cautions that “Dialectic is not a machine producing automatic results. It implies the certitude of human responsibility and therefore a freedom of choice and decision.” [9].

Generalizing from the above four notions of dialectics in a manner relevant to this book, we can identify four factors for a challenge:

1. Coexistence of thesis and antithesis
2. Change and negatives derive challenges
3. Synthesis is an emerging phenomenon
4. Noise.

While noise was not an explicit topic in the above discussions, it needs to be induced. Given the many contradictions that exist in the world with no potential to influence a challenge, they can inhibit the emergence of challenges when attempting to automate the process. Therefore, they should be filtered out. Given the nature of this noise, it is best suited for humans to filter them out than automation.

The above does not necessarily offer a solution to how we can model, design, and create a challenge, but it certainly offers cautionary features for which we need to account for when discussing automation. As a principle, this book does not claim that we can automate the concept of a challenge; in fact, this is precisely why we will dismiss the concept of automating the CRT process.

CRT as a creative process, needs the “transformative moments” through which synthesis is formed in the mind of a human, where it is not planned, nor can it be explained deductively in terms of the data or the logic used within a piece of software. These transformative moments, where ideas emerge and only a vague argument or a correlate of logic can be given as a justification.

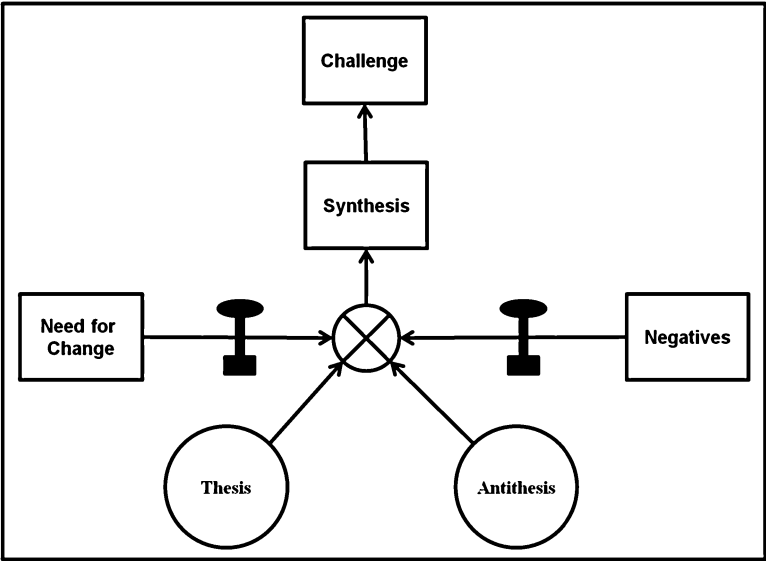


Fig. 2.13 Synthesizing Ellul and Mendoza opinions on a challenge

In this book, automation is discussed as a means to deriving a process of challenging. The advanced computational models we will discuss in this book are merely tools that can be used in a computer-in-the-loop CRT exercise.

Figure 2.13 synthesizes the concepts discussed by Ellul to form a basic structure for a challenge. This structure can be seen as an information processing lend on the concept of a challenge. Nevertheless, the views of Ellul and followers do not structure the concept of a challenge sufficiently for our computational purposes. We will attempt to do this. This will not to be an attempt to claim (erroneously) that there is one and only one model through which to structure the concept of a challenge, but to lead to the creation of more models on how to structure the concept of a challenge in future research.

This model must surpass the classical conceptual models offered in fields such as dialectics and the indirect and unstructured manipulation of the topic in areas such as systems thinking; mainly because our model should also allow for automation to occur. Therefore, while we need to seed it with these areas of the academic literature, we need to ground it in concepts that are computable; processes that can be automated; and methodologies that can be objectively assessed.

We will not present the model directly, rather, we will evolve it so that the reader can subscribe to the logic behind the model. We will begin with the simplest ones, then proceed steadily to a more meaningful, realistic and functional model.

We will begin with two simplistic representations that capture the concept of a challenge from different perspectives. Figure 2.14 depicts the first conceptual representation of a challenge. It is an intuitive model that subscribes to a common

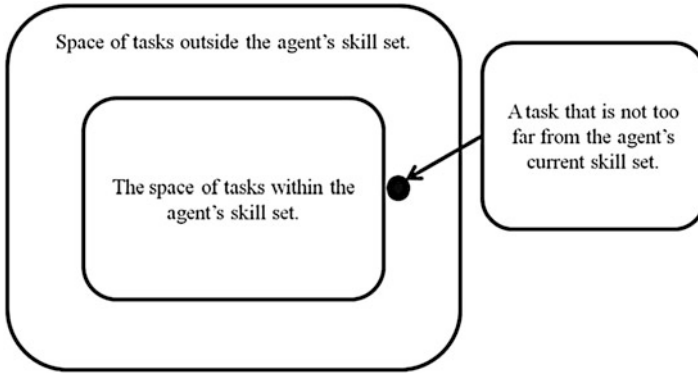


Fig. 2.14 A conceptual diagram of the concept of challenge

intuition of a challenge. The totality of the skills an agent possesses represents the set of tasks the agent is capable of performing. To challenge this agent is to find such task that the agent cannot perform because the agent lacks certain skills, while at the same time, this task is very close to what the agent can currently do.

For example, we ask a child to multiply three by three knowing that the child has learned how to add and knows the basics of the concept of multiplication, for example, knowing how to multiply two by two. However, the child is unable to multiply three by three because the child has not been exposed to sufficient examples to generalize the concept of multiplication to arbitrary multiplication of any two numbers. Nevertheless, the child was able to generalize the concept of addition to arbitrary numbers, and understands the basics of multiplication in the simple example of multiplying two by two. The child has all the skills to multiply three by three, except one skill: the know-how to generalize that multiplication is a recursive addition. Whether or not this extra step is simple enough or too hard for the child will depend on the child's cognitive resources and skills.

Likewise, we can teach a person how linear regression works then challenge them by giving them a simple nonlinear example that requires a simple transformation to make it linear. The person needs to synthesize their knowledge to solve the example in a manner in which they have no experience. Even if they fail, once the solution is explained to them, they see no problem in understanding it. This is the point at which we hear exclamations such as "Ah, I see, this now sounds so obvious, it just did not cross my mind the first time I attempted to solve this problem."

The above example demonstrates an important point that many people may find counterintuitive, that is, a challenge can only synthesize existing knowledge, it cannot introduce new axiomatic knowledge. Now is a good time to differentiate between these two types of knowledge.

We will argue that there are two broad categories of knowledge an agent can have: axiomatic and derivable (learned) knowledge. Axiomatic knowledge can only be gained through direct exposition to certain facts, processes, and tasks. Similar

to mathematics, once we believe in the axioms, theorems can be derived from the axioms, both deductively or inductively. To develop a new type of calculus, it is not sufficient to study and practice calculus, we need different types of knowledge to understand what it means to develop a new calculus in the first place.

Similarly, people who studied humanities may be very creative when writing a story or analyzing a conversation. However, if they have never studied mathematics, no challenge can synthesize their existing knowledge into a new type of knowledge that enables them to understand mathematics. The distance between the two spaces of knowledge is large. The same result will ensue by asking a mathematician to understand Shakespeare if they have not been exposed to literature before or by asking a person who has recently begun to study a language to understand complex jokes in that language; we know this is difficult because a joke does not just play with words in a language, it also relies on cultural elements that the person may not have gained this type of axiomatic knowledge of this particular culture.

This is not to say that a challenge does not produce new knowledge; on the contrary, if it does not, then it is not a challenge. Instead, a challenge can only move us from one place to a place close by; thus, the knowledge the challenge produces may impress us but it must come from within a space that is sufficiently close to the space of the old knowledge. This knowledge can be “transformative”—as Ellul indicated with “transformative moments”—in the sense that it is a non-linear synthesis of existing knowledge. Because of non-linearity, it is hard to explain it deductively from existing knowledge. The agent may perceive that it is new axiomatic knowledge, but the agent would feel also that it is not too difficult and that it can vaguely be associated with what they already know.

Recasting the previous conceptual diagram of a challenge in a different form, the skills of an agent would influence the agent’s behavior. Figure 2.15 depicts this process by conceptualizing the space of possible behaviors an agent can express.

The model assumes that we wish to challenge a thinking entity, let us refer to this entity as an agent. Similar to the theory of challenge in the field of education, our aim is to push further this agent to acquire skills and knowledge beyond those it currently possesses.

Figure 2.15 offers a complementary perspective on a challenge when the aim is to challenge the behavior of an agent or a system; the aim is to encourage the system to express a behavior that is outside the boundary of its normal behavior. For example, challenging a passive person to take a more proactive attitude should not be considered a process that will magically transform this person into a proactive person overnight. It is extremely unlikely that such a transformation will occur so rapidly simply because being proactive requires many skills to be acquired, including thinking and communication skills.

Within this space resides a subspace of the behaviors the agent currently expresses, which we assume in this example to represent the space of passive behaviors. To expand the behavior subspace of the agent to include proactive behaviors, the small dark circle represents the closest subspace that features proactive behaviors but is not too far away from the agent’s current subspace of behaviors.

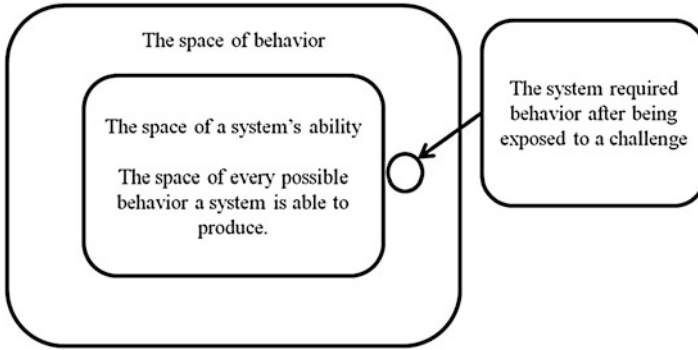


Fig. 2.15 A conceptual diagram of the concept of challenge

To achieve the intended effect from a challenge, the agent must be engaged during this process, that is, the agent should not find the challenge process too boring or too difficult, but instead, stimulating and motivating. The challenge needs to stimulate the agent so that a new behavior is expressed by the agent. Therefore, to ensure that the challenge is effective in achieving the desired effect, its design needs to be agent-centric to connect agent's skills with agent's motives. To this end, and before progressing any further, we must pause to explain what we mean with behavior in this context.

2.4.4 Challenging Technologies, Concepts and Plans

Figure 2.16 expands this discussion beyond the use of the concept of a challenge to expand the skill set or behavioral subspace of an agent, to testing and evaluating algorithms and systems. This new example will allow us to dig deeper in an easy to understand context. In Fig. 2.16, we assume a computer network. In this environment, A represents the space of all behaviors or all possible traffic that goes through this network. Some of this traffic will constitute anomalies and is depicted by the subspaces B and D . The difference is that we know of the existence of B but we do not know of the existence of D because of our bounded rationality, limited knowledge or any other reason that would prohibit our thinking from knowing about the types of anomalies hidden in D .

We can assume an algorithm that is able to detect anomalies. This algorithm may be able to detect anomalies in subspace C , which is a subset of B . A classical test and evaluation method such as stress testing to evaluate this algorithm will very likely end up with the subspace $B - C$. This is because the bias that exists in our design of these stress-testing methods is (subconsciously) based on our knowledge of B .

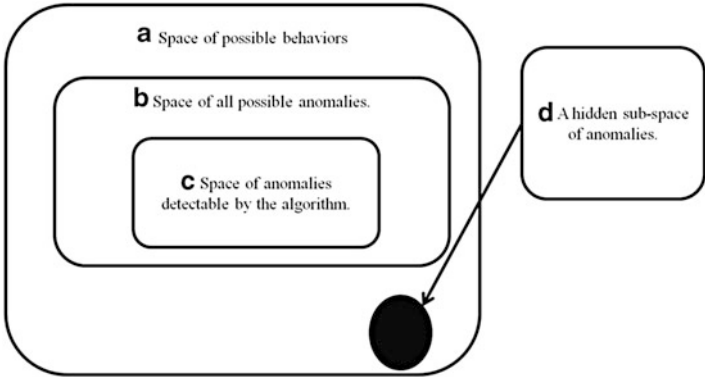


Fig. 2.16 A conceptual diagram of the concept of challenge

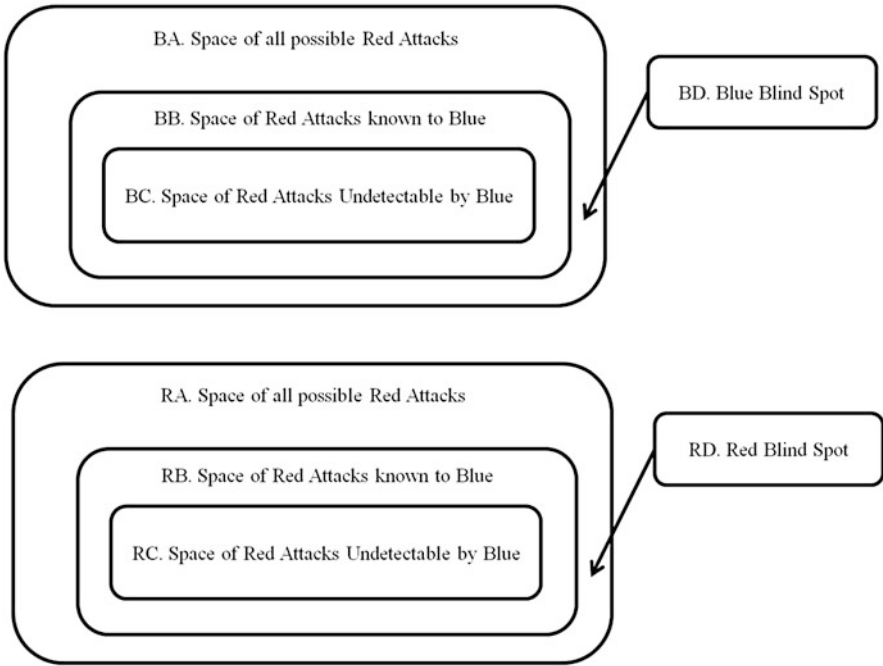


Fig. 2.17 A conceptual diagram of the concept of challenge

Methods designed based on the concept of a deliberate challenge should discover that in addition to $B - C$ exists the subspace D . Notice that D is sufficiently close enough to B , but resides outside our sphere of bounded rationality. It is reachable through CRT.

Thus far, we have ignored the fact that a challenge requires at least two agents: one challenges the other. Figure 2.17 demonstrates this by depicting the two design

spaces of a challenge for both teams. In the top space, blue searches the space of red attacks. While there is a space of red attacks that is known to blue, there is a subspace within this space where blue knows that if red attacks come from this subspace, blue can't detect them. This is the subspace where blue is aware and conscious of its own vulnerability.

There is also a subspace in the space of all possible attacks by red, where blue is unaware of it. Thus, this subspace represents the blind spot for blue. The same analysis can be done on the red side in the bottom diagram.

2.5 From the Analytics of Risk and Challenge to Computational Red Teaming

Sofar, the discussion introduced many concepts that underpin the risk and challenge analytics areas. It is time to synthesis these introductory materials into computational forms. The discussion will start with the first formal perspective on CRT and how it relates to the analytics of risk and challenge. This will be followed by a more focused discussion that synthesizes the introductory materials into a coherent form for each of the cornerstones of CRT.

2.5.1 From Sensors to Effectors

The grand challenge in computational red teaming is to seamlessly transform sensorial information to effectors that create the right set of effects. This grand challenge is depicted in its simplest form in Fig. 2.18.

This figure is too generic and goes beyond the realm of CRT. One can say that it is a picture that captures the generic objective of autonomous systems: how to transform sensed information into an agent's desired effects through the design of effectors that influence and shape the environment.

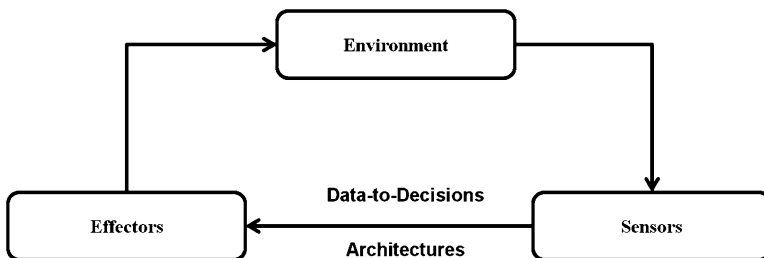


Fig. 2.18 Transforming sensorial information, from sensors, to effects, through effectors, cycle

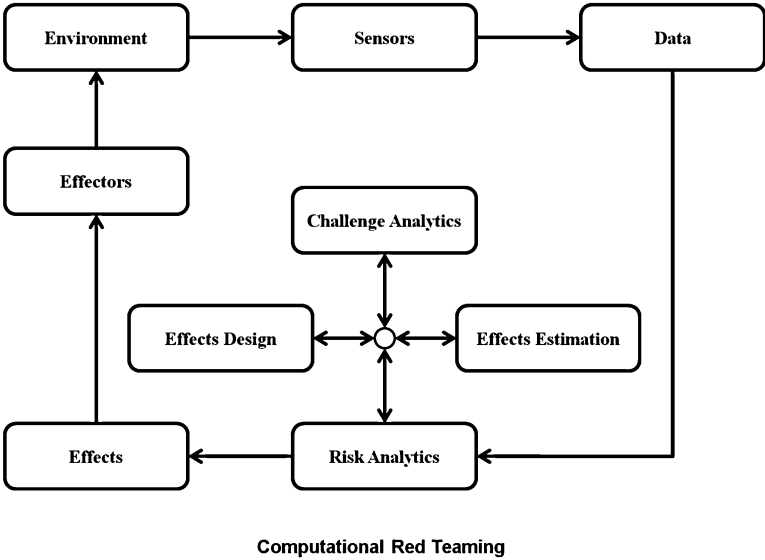


Fig. 2.19 Cognitive–cyber–symbiosis of the CRT-based sensors-to-effectors architecture

Zooming in on this pictorial representation of an autonomous system from a CRT perspective, Fig. 2.19 expands the picture with the cornerstones of a CRT system. A CRT system should not be understood as a simple computer program, but a system of systems to conduct a CRT exercise; some components can be software-based, while others are human-based. These components interact together to form a CRT system. We will call this process Cognitive-Cyber-Symbiosis (CoCyS)—pronounced as “Cookies”—to emphasize that this system is not based on the mere integration of different components, but on blending and seamlessly fusing these components to form a single computational CRT machine in a fluid manner.

Definition 2.19. Cognitive-Cyber-Symbiosis (CoCyS) is an environment whereby human thinking, mental processes and indicators, and the Cyber space are blended to improve the effectiveness of decision making.

The first cornerstone of a CRT system, as shown in Fig. 2.19, is risk analytics. This agent aims at analyzing how uncertainty impacts the system’s objectives. More details on risk analytics are offered in Sect. 2.5.3. The second cornerstone of CRT, Challenge Analytics agent, is linked to the risk analytics agent. Challenges are discovered and formed using the Observe-Project-Counteract (OPC) agent architecture discussed in Sect. 2.5.4. In essence, risk analytics in CRT analyzes risk by challenging its own decision making and thinking process as well as the external environment including competitors and other agents.

The four agents: risk analytics, challenge analytics, effects estimation and effects design, represent the four primary thinking processes that negotiate with each other

continuously to red team. These agents should be seen as an attempt to structure CRT, not an attempt to define four independent systems that needs to come together to form CRT.

The four agents described in Fig. 2.19 will share computational needs, thus, it is not advisable to duplicate the software infrastructure. Duplicating software components in a system can create many problems. First, it increases the acquisition cost of the system, simply because the same functionalities get bought from different vendors. Second, it increases the maintenance cost of the system. Third, and most importantly, over time, the CRT system becomes a can of worm: unplanned redundant functionalities and data that no one can understand the assumptions spread across a system that is supposedly designed to challenge assumptions, etc, in other systems.

To overcome the problems mentioned above, Fig. 2.20 shows the service-oriented architecture (SOA) to generically define—in a structured and not too decentralized manner—the high-level services required for CRT. An SOA is a computer architecture, whereby functionalities are defined as services without any central control mechanism. These services communicate to each other through a platform, known as the service bus, that enables services to define, discover, and use other services in the system.

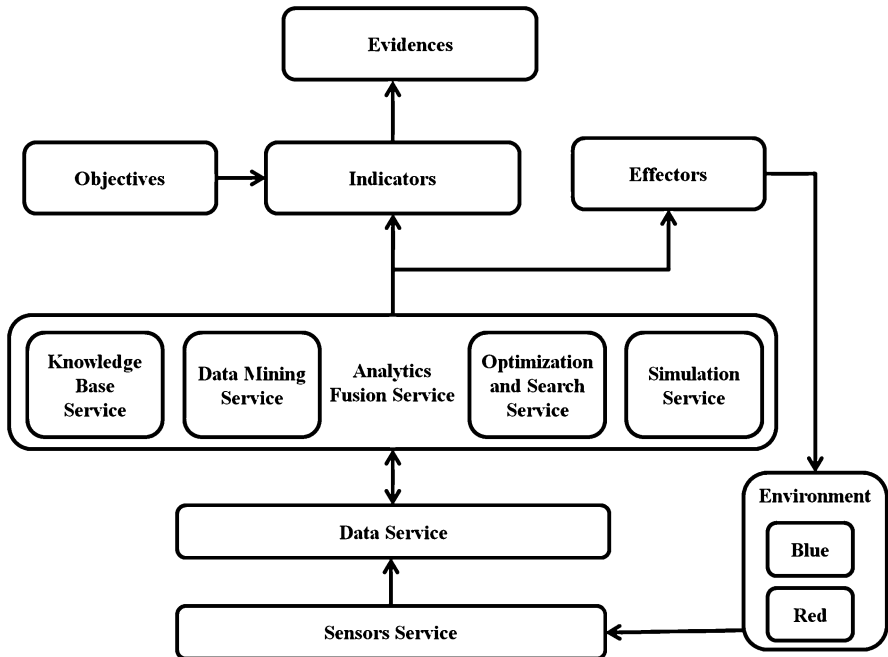


Fig. 2.20 Service-oriented architecture for computational red teaming

The technical details for implementing an SOA is beyond the scope of this book, but the concept of SOA is simple enough to be understood on this level of abstraction. SOA can be implemented using web-services; which relies on the internet as the backbone for the service bus.

Figure 2.20 shows one view of the SOA for CRT, which connects sensors to effectors. It also emphasizes that the system internally measures indicators for the success of achieving the objectives; thus, providing evidence-based decision making approach. The risk analytics component has a number of services, including optimization and simulation services. The role of these technologies will be discussed in the following chapter.

Both challenge analytics and risk analytics rely on three technologies: simulation, optimization and data mining, similar to risk analytics. These technologies will be discussed in more details in Chap. 2, and an example to illustrate how they need to work together for CRT purposes is given in Sect. 3.1.2.

2.5.2 The Cornerstones of Computational-Red-Teaming

Risk analytics and Challenge analytics are the two cornerstones of a CRT system. Figure 2.21 depicts this relationship by factoring risk into its two components: uncertainty and objectives. Together, the objectives of the organization and the uncertainty surrounding the decision making process constitute risk. The challenge analytics component aims at designing challenges for uncertainty, constraints and objectives. This point will be elaborated on in Sect. 2.5.4.

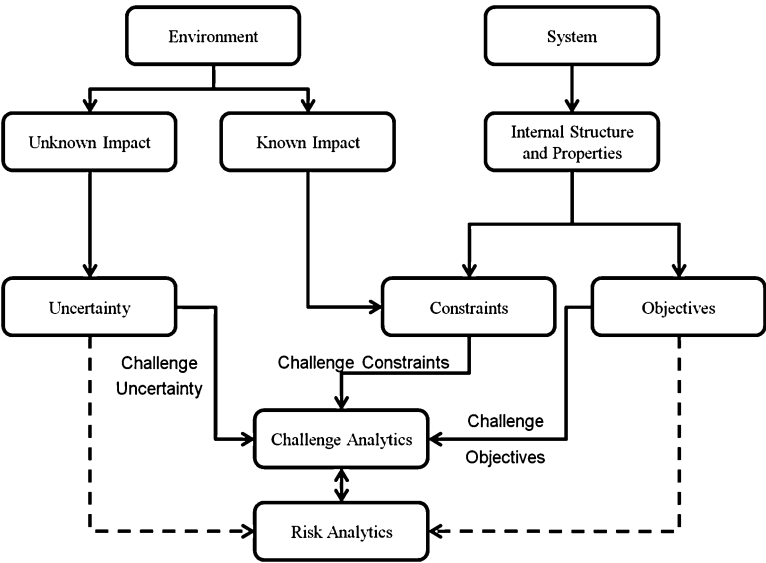


Fig. 2.21 The cornerstones of computational red teaming

2.5.3 Risk Analytics

Risk analytics is the process of deconstructing system-level risk in an organization to its constituent parts, assess the system-level risk, and evaluate the impact of possible courses of actions on organizational risk spatially and temporally.

The difference between risk analytics and risk management is that, the former emphasizes the thinking process and tools, while the latter emphasizes the process.

Risk analytics (see Fig. 2.20) begins with transforming input data into effects' indicators. The CRT system is designed for a purpose, which is defined with a set of effects that need to be generated, and therefore monitored. Before acting on any data that has been generated, either from the environment or internally within the system, the system needs to be able to measure some indicators of the effects it aims to generate. This measuring process acts as a reinforcement mechanism for the system, enabling the system to measure deviations from intended effects and correcting its actions accordingly. The Effects Estimation agent is responsible for continuous monitoring of the environment to estimate effects' indicators.

The output of risk analytics is a set of effects. The risk analytics agent needs to take into account the environment, including the properties of the agents in that environment and how effects should be shaped and framed to influence these agents. This is the role of the Effects Design agent.

From the definition of risk as the impact of uncertainty on objectives, the two cornerstones of risk analytics is to analyze uncertainty and objectives to synthesize a risk picture. This risk picture needs to be challenged to manage risk. The challenge analytics component is responsible for finding actions that can have a positive and/or negative impact on organizational objectives.

2.5.4 Challenge Analytics Using the Observe-Project-Counteract Architecture

The process for challenge analytics consists of two steps:

Estimate Boundaries: In this step, the boundary constraints are discovered. Section 3.1.2 will provide an example on how a challenge can be computationally discovered. In its basic form, the process for blue (red) works by estimating the boundary constraints of red (blue) capabilities.

Probes: Once boundaries become known, the probing step attempts to design search mechanisms to generate points just across the boundaries. When a probe crosses the boundaries with a small distance, it becomes a challenge.

As we discussed before, risk is made up of uncertainty and objectives. Challenge analytics, as shown in Fig. 2.21, attempts to design ways to challenge uncertainty, constraints, and objectives so that CRT can challenge risk.

The challenge analytics of uncertainty aims at estimating the boundary constraints of uncertainty; that is, instead of enumerating what may happen, it discovers the boundary on what may happen. Uncertainty may be misunderstood as an unbounded space of possibilities. However, life is evolutionary, not revolutionary. What bounds uncertainty is plausibility. Anything can happen, but nothing happens in vacuum.

Plausibility bounds uncertainty and plausibility depends on know-how (skills).

The prices of shares in the stock market can rise very quickly, but we can estimate a boundary on how far they can rise. It is possible to estimate multiple boundaries with different levels of confidence. If blue can estimate the bounds on red's uncertainty, blue can design strategies to challenge red by creating uncertainties outside these bounds.

Similarly, challenge analytics need to challenge objectives. We discussed that classical decision making assumes that objectives are mostly defined and fixed. However, objectives in CRT are controllable elements that can be reshaped. One way for blue to influence red is to reshape red's objectives. Challenge analytics can help blue to estimate the boundary conditions on red's objectives so that blue can challenge red by aiming to reshape red's objectives. This reshaping process can be done by changing these boundaries, moving them in different directions.

To illustrate a simple example using classical linear programming, assume that red aims to maximize profit, where the profit objective function is formulated as follows:

$$\uparrow 2 \times x + 3 \times y$$

with x and y representing two different types of effects that red wishes to generate. For blue to challenge these objectives, blue needs to analyze two different boundaries: the boundaries on the coefficients, and the boundaries on the structure.

The boundaries on the coefficients is to estimate how far the two coefficients of 2 and 3 for x and y can change, respectively. However, some gains achieved by red are influenced by blue. These coefficients represent red's gain from each type of effects. In essence, they represent how red values these effects. As such, to challenge these coefficients is to understand the boundary constraints on them; that is, for example, the coefficient of x may change between 1 and 5 based on a number of factors. Blue can then design a strategy to influence these factors so that this coefficient changes in the direction desired by Blue.

The boundaries on the structure aims at estimating the constraints on the effect space for red. In other words, can we introduce a third variable z to this equation that is more beneficial for us? These structural boundaries are very effective tools.

Fields such as Mechanism Design and Game Theory can assist in discovering this third dimension, although we will avoid discussing Mechanism Designs in this book because most work in this domain falls in the same classical trap of game theory, which assumes (1) rational agents and (2) agents are self-aware of the value of any alternative (i.e. when an agent is faced with an alternative, the agent has an internal value representing the maximum value the agent would be willing to pay for that alternative). The advantages of CRT is that, it does not have such restrictive and unrealistic assumptions. For example, what would be the maximum price you would be willing to pay to save your life? In essence, the question also means, how far can you go to save your own life? Mechanism design assumes that each agent knows the answer to this question precisely!

The third element that CRT can challenge is the constraints on the other team. Constraints normally exist for two reasons; either the structure and properties of the system are inhibiting the system from expressing certain behaviors, or the environment is doing so. Constraints from the environment are forces impacting the system in a similar way to uncertainties. The primary difference between an environmental constraint and uncertainties is that the former are certain forces, while the latter are uncertain. For example, weather conditions are environmental conditions impacting a flight. When weather conditions are known, we can take them as constraints when designing an optimal flight path. When weather conditions are not known, they become uncertainties that a flight path needs to be evaluated against a range of possible weather conditions. In classical optimization, the two concepts can be combined in the form of a stochastic constraint.

Most of the time challenge analytics is concerned with designing counteractions to challenge the other team. This design process for blue(red) will require mechanisms to estimate boundary conditions for red(blue) constraints, uncertainties and objectives, designing actions outside these boundaries, projecting the impact of these actions in the future, and selecting the most appropriate counteraction for blue(red) in response to red(blue) actions.

As will be illustrated in Sect. 3.1.2, challenge analytics rely on three technologies: simulation, optimization and data mining, similar to risk analytics.

Computationally, challenge analytics requires a proactive architecture that can support proactive generation of counteractions. One possible realizations of this architecture is the following Observe-Project-Counteract agent architecture. This architecture has three components as follows:

Observe: In the first stage, each team needs to observe the other team by continuously sensing information, extracting behavioral patterns, and assessing their skills (assessing boundary constraints).

Project: In the second stage, the creation of a model of how the other team acts is required, so that each team can use this model to estimate their actions in the future, and evaluate the impact of one team's actions on the other team. In the debate, if we can estimate through observations what the other team knows, we can equally estimate their response to our future questions.

Counteract: In the third stage, counter-strategies are designed to counteract what the other team intends to do. The ability to observe and project the other team's behavior into the future provides a team with the means to evaluate its counter-strategies.

Many variations can be created from this architecture by replacing the word "observe" with "sense," and replacing the word "project" with "anticipate," "predict," or "estimate." Clearly, each of these words has a slightly different meaning, and appropriate use will depend on the context.

For example, the difference between observe and sense is that "observe" reaches beyond a basic level of sensing. Observing requires intentional sensing of information and making sense of this information.

Similarly, microscopic differences can be defined between the words project, estimate, predict and anticipate. A random guess of where blue is going in the absence of any information is a basic type of prediction. Therefore, blue can predict based on its own beliefs, without the need for information on red. Establishing proper confidence intervals around this prediction will move us from the realm of prediction to the realm of estimation. Anticipation increases the complexity even further by using future state information to define the current state of the agent. Projection is the wider concept, whereby any form of prediction, estimation or anticipation is considered a form of mapping between existing states to future ones.

The word "counteract" is emphasized instead of the word "act" because the emphasis of red is not to produce an action independent of blue. One can act out of one's own interest or even subconsciously. However, counteraction is a function of an opponent action; it is a deliberate response that requires proper logic to be in place to undo deliberately the effects of the opponent's action.

The critical distinction between a counteraction and a response resides within the clause "to undo deliberately the effects of the opponent's action." A counteraction is not a simple or reactive response, but a response that is designed with the effect of the competitor's action in mind. It is a response designed to ensure that the effect of the competitor's action is not materialized.

Deliberate actions are centered on the objective of an agent. When the level of this objective relies also on the actions of the opponents, the agent's action becomes a counteraction.

References

1. Abbass, H.A., Petraki, E.: The causes for no causation: a computational perspective. *Inf. Knowl. Syst. Manag.* **10**(1), 51–74 (2011)
2. Einstein, A., Infeld, L.: *The Evolution of Physics*. Simon and Shuster, New York (1938)
3. Ellestad, M.H.: Stress testing: principles and practice. *J. Occup. Environ. Med.* **28**(11), 1142–1144 (1986)
4. Gaidow, S., Boey, S., Egudo, R.: A review of the capability options development and analysis system and the role of risk management. Technical Report DSTO-GD-0473, DSTO (2006)
5. Gilbert, T.F.: *Human Competence: Engineering Worthy Performance*. Wiley, Chichester (2007)

6. Grimshaw, M., Lindley, C.A., Nacke, L.: Sound and immersion in the first-person shooter: mixed measurement of the player's sonic experience. In: *Proceedings of Audio Mostly Conference* (2008)
7. <http://www.thefreedictionary.com/>. Accessed 1 Feb 2014
8. ISO: ISO 31000:2009, Risk Management - Principles and Guidelines (2009)
9. Mendoza, S.: From a theory of certainty to a theory of challenge: ethnography of an intercultural communication class. *Intercult. Commun. Stud.* **14**, 82–99 (2005)
10. Sanford, N.: *Self and society: social change and individual development*. Transaction Publishers, Brunswick (2006)
11. Sawah, S.E., Abbass, H.A., Sarker, R.: Risk in interdependent systems: a framework for analysis and mitigation through orchestrated adaptation. Technical Report TR-ALAR-200611013, University of New South Wales (2006)

Computational Red Teaming
Risk Analytics of Big-Data-to-Decisions Intelligent
Systems

Abbass, H.A.

2015, XXIII, 218 p. 61 illus., 15 illus. in color., Hardcover

ISBN: 978-3-319-08280-6