How do adaptive immune systems control pathogens while avoiding autoimmunity?

Carl T. Bergstrom¹ and Rustom Antia²

¹Department of Biology, University of Washington, Seattle, WA 98115, USA
²Department of Biology, Emory University, Atlanta, GA 30329, USA

Immune systems face a daunting control challenge. On the one hand, they need to minimize damage from pathogens, without wasting energy and resources, but on the other must avoid initiating or perpetuating autoimmune responses. Finally, because pathogens interfere with immune function, immune systems must be robust against sabotage. We describe here how these challenges are met by two immune systems, the intracellular RNA interference system and the vertebrate CD8 T-cell response. We extrapolate from these two systems to propose principles for strategically robust control.

Introduction
Physiologists consider the purpose, function or goal of a biological structure when trying to understand how that structure works. Immunologists do the same thing. The goal of any immune system is to protect against pathogens and these systems have therefore evolved to increase the fitness of the organism by reducing the damage caused by such organisms [1], ideally without wasting energy and resources [2]. To use this functional approach successfully, one must account for the tradeoffs and constraints that organisms face. Here, we focus on two that have been instrumental in immune evolution: (i) Autoimmunity: immune systems need to minimize the risk of autoimmunity. A single autoimmune mistake is potentially lethal, if directed against essential components of the body; and (ii) Pathogen subversion: immune systems must be strategically robust. They need to work in ways that rapidly evolving pathogens cannot exploit, subvert, or sabotage.

Many systems, including the immune system, must be robust: they need to operate in a range of background conditions, function in the presence of noise and despite variation in internal structure, and keep working even if multiple internal components fail. Mechanisms of robustness have been studied extensively in engineering [3] and biology, from the biochemical level [4] to that of the ecosystem ([5]; see also [6,7]). Systems that have to deal with internal subversion must go one step further and be strategically robust: that is, they need to function properly despite efforts to sabotage their workings.

The distinction between robustness and strategic robustness becomes clear through analogy. A robust computer circuit would function effectively even if a few resistors burned out at random. A strategically robust computer circuit would function even if a disgruntled technician tried to sabotage the machine by removing precisely those resistors that were most crucial. As we go from robustness to strategic robustness, we go from a simple optimization problem to a game-theoretic one, in which antagonists each try to maximize their own payoffs at the possible expense of the others. Thus, the task of implementing strategic robustness is as much in the spirit of the mechanism design problem from economics [8], in which the designer aims to set up the rules of the game so as to make multiple self-interested players behave as the designer wishes, as it is in the spirit of control theoretic approaches from engineering.

The task of an immune system (Figure 1) is especially difficult because efforts that meet one challenge often compromise efforts to meet another. To avoid autoimmunity, immune systems must have ways of terminating accidental self-directed responses; however, these ‘shutdown’ pathways can be strategic vulnerabilities. Pathogens can and do evolve to exploit the mechanisms that immune systems use for self-regulation [9].

Here, we explore the ways that immune systems deal with the challenges of strategic robustness and autoimmune avoidance. We focus on two principle examples: the intracellular RNA interference (RNAi) pathway that many non-vertebrates use to combat viruses (Box 1), and the specific immune response that vertebrates use to deal with viruses, bacteria and other microparasites (Box 2).

RNAi, an intracellular immune system in non-vertebrate eukaryotes
RNAi is a system of post-transcriptional gene silencing that is broadly conserved across eukaryotes; it appears to have evolved as a form of adaptive immunity to prevent viruses from replicating within infected cells, by targeting foreign nucleic acids. Although the RNAi system has been co-opted for gene regulation in vertebrates, it retains immune function in organisms as diverse as unicellular eukaryotes (e.g. yeast), invertebrates and plants [10–12].
Similar to the vertebrate adaptive immune system, RNAi mounts and amplifies a highly specific response against pathogens [13] (Box 1). When directed toward legitimate targets, such as viral mRNAs, these responses require careful regulation to avoid energetically costly runaway amplification. However, when directed toward illegitimate targets, they rapidly turn small mistakes into big problems. Thus, if a cell accidentally produces double-stranded RNA (dsRNA) corresponding to its own genes [14], this dsRNA can induce a massive amplification of a self-directed RNAi response and thereby silence the corresponding self-genome. An accidental reaction in a single cell can silence gene expression throughout an entire tissue, because RNAi reactions spread systemically in many taxa [15]. To avoid these problems, the RNAi pathway guards against mistaken responses, and deploys several mechanisms that limit the damage when mistaken responses do occur.

**Unidirectional amplification**

As with all known nucleic acid polymerases, the RNA-directed RNA polymerase used in RNAi operates in one direction, from 5′ to 3′ along the nascent RNA strand (but see [15]). This unidirectionality guards against runaway amplification. Each short interfering RNA (siRNA) is amplified into a dsRNA that is, on average, shorter than the dsRNA from which it was produced. After repeated rounds of amplification, the remaining siRNAs correspond primarily to downstream portions of the original dsRNA and, when these bind to mRNAs, the polymerase has nothing left to copy [16]. The RNAi pathway thus acts as a self-limiting amplifier: a silencing reaction will persist only if it receives ongoing dsRNA input, which can occur when a viral infection has generated the dsRNA, but is unlikely following accidental dsRNA production.

**Spatial discrimination**

The RNAi pathway uses spatial cues to distinguish between self and non-self RNAs. Typically, self-RNAs originate in the nucleus, whereas viral RNAs enter through the cytoplasm. Knight and Bass [17] hypothesize that a group of RNA-editing enzymes exploit this distinction to prevent self-directed reactions. The adenosine deaminases that act on RNA (ADARs) are present in the nucleus but not in the cytoplasm [17], and inhibit RNA interference by inducing structural changes in dsRNA [18]. By contrast, parts of the RNAi cycle occur in the cytoplasm. Thus, when a dsRNA originates by

---

**Box 1. RNA interference as an intracellular immune system**

RNA interference (RNAi) operates within individual cells to deter viral replication. Figure 1 provides an schematic overview of this pathway.

In step one (Figure 1a), the RNAi pathway uses dsRNA as an indicator of non-self. Long segments of dsRNA should not be present in properly functioning eukaryotic cells, but they do occur at some replicative stage of many RNA viruses. In step two (Figure 1b), the dsRNA is degraded into ~22 nucleotide fragments by the Dicer enzyme. In step three (Figure 1c), these short fragments, or siRNAs, are stabilized by a set of proteins known as the RISC complex. Half of the siRNAs are complementary to the viral mRNAs (mRNAs), and thus serve as specific templates for identifying those mRNAs. In steps four and five (Figure 1 d,e), once the appropriate target mRNAs are located, they are either degraded or, alternatively, a host-encoded RNA-directed RNA polymerase (RDRP) synthesizes additional specific templates, thereby amplifying the initial response.

---

**Figure 1.** Adaptive immune systems. An effective immune system must prevent damage from infection while avoiding subversion by pathogens and autoimmune mistakes. As indicated by the dashed arrow, efforts to avoid autoimmunity might compromise efforts to avoid subversion, and vice versa.
Box 2. Clonal selection and expansion of the vertebrate immune response

The adaptive immune response of vertebrates works by clonal selection. Independent of exposure to an antigen or pathogen, the immune system generates a repertoire of immune cell lineages or clones (labeled 1–8 in Figure I), each encoding a receptor with a predetermined shape and specificity. The human immune system creates in excess of 10⁶ different clones. As a first approximation, those that react with self-antigens (numbers 3, 5, and 8 in Figure I) are deleted shortly after they mature. When the individual is infected with a pathogen, those clones that are specific for the pathogen (number 2 in Figure I) will proliferate, producing a pathogen-specific immune cell population that is large enough to control that pathogen. This process is known as clonal expansion. After the pathogen is cleared, some of the pathogen-specific immune cells survive and confer immune memory.

![Figure I](image)

accident in the nucleus, it is converted to an inactive form by the ADARs and it does not trigger the RNAi pathway; however, when a dsRNA originates in the cytoplasm, it triggers the RNAi pathway before the ADARs can modify its structure.

Thresholds for amplification

The RNAi pathway exploits dsRNA copy number differences to avoid amplifying mistakes. Viral dsRNAs will often be present in higher copy numbers relative to accidentally produced dsRNAs. Thus, enzymatic activity that antagonizes the RNAi amplification cycle can disproportionately eliminate accidental reactions. Although breaking down the 'effector' molecule might be energetically inefficient, it reduces the risk of self-directed reactions. The recently discovered siRNAase ERI-1 [19] might be doing this in breaking down siRNAs. Mathematical models suggest that ERI-1 induces a threshold in the system: above a crucial concentration of siRNA, a silencing reaction will begin, but below this, the RNAi response will not be amplified [20]. If viral infection is more likely to cross this threshold than is accidental production of dsRNA corresponding to self-genes, ERI-1 could reduce the probability that a mistake leads to a silencing reaction, without appreciably altering the probability that a properly targeted response does so.

Overcoming subversion by pathogens

These three safeguards work in concert, each reducing the damage associated with accidental reactions in a different way. ADARs reduce the probability that a mistake is made. When a mistake is made, ERI-1 reduces the probability that the amplification cycle begins. If the amplification cycle is initiated under inappropriate
circumstances, unidirectional amplification shuts it down before it continues for too long. But what about avoiding pathogen subversion? Viruses and viroids have evolved numerous artifacts to evade [21], disable [22–25] or redirect [26] the RNAi response. Can pathogens also exploit the safeguards described above?

Some of the autoimmunity safeguards appear hard to subvert. Unidirectional amplification limits the magnitude of an immune response, but it should be difficult for pathogens to abuse. It enables the silencing reaction to shut down over time of its own accord, rather than in response to an external signal. Thus, it terminates reactions without the strategic vulnerability of an explicit off-switch.

The ADAR mechanism is also hard to subvert. To infect a host cell, viruses enter through the cytoplasm, where they trigger the RNAi pathway, rather than through the nucleus where they might be protected by ADAR. Furthermore, most RNA viruses replicate, and thus produce dsRNA genomes or replicative intermediates, in the cytoplasm. Segmented negative-stranded RNA viruses are the main exception as they replicate in the nucleus and, thus, could use the ADARs to help them hide from the RNAi pathway. We predict that RNAi will be comparatively less effective against those few segmented negative-stranded RNA viruses that attack plants or other taxa with active RNAi immune defense.

The ERI-1 mechanism appears easier to subvert. A pathogen could upregulate ERI-1 expression and thereby slip below the threshold density of siRNA necessary to induce a silencing reaction. But this might be difficult for the pathogen to do, because of timing; a pathogen would have to enter the cell and upregulate ERI-1 before being detected by the RNAi system. Thus, we expect that the viruses most likely to exploit this potential vulnerability will be those whose life cycles involve extensive gene expression before replication and production of dsRNA.

### The adaptive immune response of vertebrates

Clonal selection theory provides a conceptual model of how vertebrates deploy a cellular immune system in which B and T cells generate specific responses to pathogens. Mathematically, the clonal selection model can be described as an ecological predator–prey system, with the pathogen as prey and the immune response as predator [27,28]. Predator–prey models of immune dynamics predict an efficient and proportionate immune response: pathogens that are easy to clear elicit smaller responses, and pathogens that are hard to clear generate larger responses. But how does the immune system rapidly clear the pathogen without excess energetic expense? Segel and colleagues [29] suggest that the adaptive immune system solves this control problem through monitoring and feedback. By monitoring the efficacy of different branches of the immune response, the immune system can direct resources to the most effective responses [30,31].

Responding efficiently is only part of the challenge; as with RNAi, the vertebrate immune system must avoid responding to self [1]. Here, we focus on how one component of the vertebrate immune response, the CD8 T cell or cytotoxic T cell (CTL) response, avoids self-directed responses. Because CTLs kill the targeted cells, a self-directed response can cause destruction of self-tissue [32] with severe consequences. For example, some forms of type 1 diabetes are caused by a CTL response that destroys islet cells in the pancreas [33].

### Clonal deletion during ontogeny

The first step in avoiding self-directed T-cell responses is to delete self-reactive cells just after they are generated. The basic logic is this: a T cell will be stimulated by antigen shortly after it is produced if its T cell receptor (TCR) has responded to a self-antigen. The cell is potentially self-reactive and is deleted in the thymus where T cells mature.

### Two signal and danger models

Such clonal deletion is useful, but it is not sufficient to prevent all self-directed responses. For instance, this process cannot remove CD8 cells that act against self-antigens that are not found in the thymus; neither can it remove CD8 cells that are specific for those self-antigens that are expressed only during later developmental stages of the adult.

The two signal [34] and danger [1] models explain how self-reactive T cells can be deleted or inactivated after they leave the thymus. Both models are based on the observation that a T cell requires two signals to proliferate. The first signal is delivered via the TCR and ensures that the T cell responds only to its specific antigen. The second signal is thought to come from the pathology or cell damage caused by infections. If the T cell receives the first (antigen-specific) signal in the absence of the second signal, it is inactivated rather than stimulated and is unable to respond thereafter. This process of inactivation, termed ‘anergy’, explains why most of the self-specific T-cells that are not deleted in the thymus do not cause future problems.

### Clonal exhaustion

Although clonal deletion in the thymus and anergy outside of the thymus prevent most potential self-directed reactions, neither can control a self-directed response once it begins. Here, a third safety mechanism, known as clonal exhaustion, becomes important. If an immune response is generated and the associated antigen does not disappear, it is likely that the immune system has targeted either a pathogen against which it is ineffective, or a self-antigen. Either way, there is potential harm and little benefit to continuing the immune response. Thus, when T cells are stimulated for a prolonged period by a specific antigen, they are down regulated by the process of clonal exhaustion [35,36].

### Working together

Similar to the mechanisms that prevent self-directed RNAi, the mechanisms of autoimmune protection work in concert. Deletion in the thymus and anergy in the periphery reduce the probability of autoimmunity, whereas clonal exhaustion reduces the damage caused...
following the inadvertent generation of an autoimmune response.

Clonal deletion in the thymus or anergy in the periphery enhance the efficiency of immune surveillance. When self-reactive cells are deleted, additional room is made for other pathogenic-specific cells. By contrast, the clonal exhaustion mechanism imposes a tradeoff between efficiency and self-reactivity. The lower the threshold for clonal exhaustion, the smaller the size of the autoimmune response following a misdirected reaction, but the higher the risk of terminating an appropriate response before clearing a pathogen that otherwise could have been eliminated.

Overcoming subversion by pathogens
Although monitoring and feedback can improve efficiency, they can also render a system vulnerable to sabotage or subterfuge. Many viral and bacterial pathogens subvert the immune system by targeting its sensing or coordinatin machinery [9,37]. Such gambits are more likely to succeed when pathogens are at high density within a host. For example, some pathogens tamper with the immune response by secreting the cytokines or cytokine inhibitors used to regulate immune responses; however, these pathogens must be present at a high enough density to produce sufficient quantities to affect the immune response.

To pre-empt such subversion, the T-cell response to a pathogen could be determined during the initial stage of the infection when the density of pathogen is low. The subsequent stages of the response should consist of proliferation in an antigen-independent manner. Recent experiments [38–40] suggest that this is what happens. Once triggered, the clonal expansion of CD8 cells continues even if the antigen signal disappears.

These findings force us as modelers to replace a decade-long series of predator–prey immune models with models that incorporate a developmental program [41–43]. Antigen-independent proliferation programs might be less efficient than are proliferation strategies that involve close and continual monitoring. But they should also be more strategically robust, because antigen-independent strategies lock in before the pathogen reaches high density and, thus, are less prone to interference.

Because the pathogen environment is highly variable and, thus, unpredictable, and because fine-tuning is precluded by the locked-in program, the process of immune cell proliferation must err on the side of caution, typically overshooting the necessary number of CD8 cells required to clear the pathogen. This is what we observe. In response to many infections, CD8 cells continue to proliferate actively after the pathogen is cleared [40]. Even when experimental manipulations reduce the magnitude of the CD8 immune response, the response still clears many infections [44].

Thus, we observe a direct tradeoff between strategic robustness and the efficiency of an immune response. Earlier commitment or longer periods of antigen-independent proliferation will confer greater robustness against subversion, but at the cost of reduced efficiency. We also expect a tradeoff between strategic robustness and the ability to avoid autoimmunity. Increased commitment brings with it a risk of amplifying a response against self-antigen, at least until the response is turned off by a mechanism such as clonal exhaustion to persistent antigen.

The vertebrate immune response also pre-empts other forms of subversion. Because cytotoxic T cells recognize viruses with the help of the major histocompatibility (MHC) receptors on the surface of infected cells, a pathogen could avoid the CTL response by knocking out a component of the MHC expression pathway and thereby down-regulating MHC expression on the surface of infected cells. The immune system specifically blocks this loophole with a class of cells called natural killer cells, which search out and destroy any host cells that try to down-regulate MHC expression.

These examples, although by no means exhaustive, provide a flavor of the evolutionary constraints confronting the adaptive immune system, and illustrate the kinds of design solutions that evolve in response.

Design principles for overcoming subversion by pathogens: strategically robust control
Whereas robust control has been studied extensively [45], we know relatively little about how to design strategically robust control circuits. Robust biological systems use multiple strategies to function effectively across a range of conditions, despite internal and external noise and variation, and component failure. Krakauer and Plotkin [46,47] review these approaches, which include redundancy, feedback control, modularity, anti-redundancy and/or purging, spatial compartmentalization, distributed processing, and extended phenotypes. Some of these design principles might also facilitate strategic robustness. Here, we highlight three of those principles.

Redundancy
Similar to many robust biological systems [48], strategically robust systems use multiple redundant pathways, in this case so that saboteurs cannot benefit by knocking out single components. This approach might be particularly valuable for systems that need to be strategically robust against evolving enemies. By deploying multiple redundant defense pathways, the host can influence the evolutionary trajectory of a pathogen population. Redundant defense mechanisms reduce the selective advantage to the pathogen of knocking out a single mechanism. This, in turn, greatly increases the difficulty of the adaptive search problem that the pathogen faces when looking for a weak point in the defenses of the host. Host effects on pathogen evolution can be explained by individual selection when pathogen evolution is rapid and occurs during the duration of a single infection.

Distributed processing
Biological processes can be efficiently coordinated by broadcast signals emitted from a central control unit; for example, the pituitary gland controls numerous metabolic, developmental and reproductive processes by emitting a suite of regulatory hormones. However, central control is dangerous in that it can be easily compromised.
Strategically robust systems will avoid central control mechanisms, lest a saboteur infiltrate the control center or spooﬁ its messages.

Anti-robustness and/or purging
Cytoxic T cells of the vertebrate adaptive immune system use an anti-robust strategy to deal with pathogens. They respond to signs of trouble by purging the damaged or infected cells, rather than by trying to stabilize these cells to help them live with the problem. This can be an effective way of dealing with a threat that otherwise can propagate, and anti-robustness at the cellular level confers robustness at the (multicellular) organismal level [46]. By contrast the RNAi mechanism, which presumably arose in unicellular organisms, confers robustness rather than anti-robustness. By analogy to the vertebrate adaptive immune system, we also predict anti-robustness mechanisms associated with RNAi in multicellular organisms; for example, persistent RNAi might trigger apoptotic pathways.

Additional features of strategically robust systems
The other robustness mechanisms in Krakauer’s taxonomy are less likely to contribute to strategic robustness. Feedback control is a particular concern. While feedback control can improve efﬁciency, such gains will come with an associated risk of subversion, because that feedback can be exploited. Thus, we expect some degree of feedback control in strategically robust systems, but less than we might ﬁnd were strategic robustness not a requirement.

Beyond the design principles discussed above, strategically robust systems will have additional features that might not be expected in regular robust systems. What might these features be? Again, biological systems whose evolution has been heavily shaped by antagonistic interactions, such RNAi and vertebrate adaptive immunity, can offer useful clues. We propose four such features.

Commitment
Strategically robust systems will collect information at times when it is unlikely to be subject to interference. They will then use this information to commit to a course of action when the integrity of collected information is suspect.

Anticipation and preemptive response
Control systems will often have a few crucial strategic vulnerabilities that, if left unguarded, could be exploited. Strategically robust systems pre-empt these sorts of gambit by the adversary, deploying patches to close off the most crucial vulnerabilities. The natural killer cells described above provide an example.

Diversity
Diversity within the host population can inﬂuence pathogen evolution. For example, the selective advantage to specializing against one host defense is reduced when other hosts rely on alternative mechanisms or forms of the same mechanism. MHC diversity is a classic example. Diversity differs from redundancy in that it arises at the level of the host population. It does not confer any individual with immediate protection in the form of backup systems, but instead reduces the strength of selection on pathogens to exploit speciﬁc weaknesses that are shared by only a subset of the population. As with redundancy, individual-level selection can favor diversity, via negative frequency dependence imposed by evolving pathogen populations [49].

Cross-validation
We conclude this list with a conjecture. We expect that strategically robust systems will require multiple sources of input before initiating actions that could be beneﬁcial to adversaries. This principle differs from basic redundancy in that redundancy uses multiple independent defenses, whereas with cross-validation, a single system depends on multiple inputs from varied sources. For example, if an immune response can be down regulated by a single signal from a single cell type, the system will be highly vulnerable to spoofing. If instead the system requires three separate classes of signals from three separate cell types, spoofing becomes more difﬁcult. If the separate signals are further interdependent (such that one operates as a checksum for the other two, for example) spoofing becomes harder still. Thus, we might expect cross-validation in the many signals that cellular immune systems use to coordinate and regulate their responses.

Conclusions
Most immunologists treat immune systems much as we treat other physiological systems: they conceptualize immune systems as evolved systems that must deal with a diverse but non-evolving set of pathogens. Here, we argue that this approach, although a step in the right direction, will not be sufﬁcient for understanding the complexity of immune function. We need to uncover the ways in which immune systems have evolved mechanisms for strategically robust control.

A conceptual framework proves its worth through the testable predictions that it generates. What predictions can we make? The need for strategic robustness poses an additional challenge in immune system evolution, and mechanisms that we have listed are likely to be deployed at the expense of suboptimal performance elsewhere. Thus, our approach makes two general predictions: (i) The design and function of immune systems will often deviate from the predictions made by optimality models, even those that explicitly address tradeoffs between sensitivity and risk of autoimmunity; and (ii) mechanisms that deviate from these predictions will tend to make the immune system more difﬁcult for rapidly evolving pathogens to subvert, sabotage, or out-run.

The challenge now is to ﬁgure out how to translate these broad predictions into concrete predictions for speciﬁc biological systems. Fortunately, there is no shortage of independently evolved exemplars. The two that we have discussed here, RNAi and vertebrate adaptive immunity, operate at different scales and arose independently in the evolutionary process. Numerous additional and evolutionarily distinct systems await similar consideration. For example, bacteria use restriction—
modification systems to identify and destroy viral DNA [50]; plants [51] and invertebrates [52] have evolved diverse mechanisms of innate immunity; social insects implement self/non-self discrimination on a colony level, using cuticle hydrocarbons as cues [53]. We anticipate that comparative evolutionary analysis across this range of systems will provide significant insights into the structure and evolution of strategically robust control of immune systems.

References
17 Knight, S.W. and Bass, B.L. (2002) The role of RNA editing by ADARs in RNAi. Mol. Cell 10, 809–817