Mechanics and Mental Change

Jon Doyle

In memoriam
Joseph Arthur Schatz, 1924–2007

Abstract Realistic human rationality departs from ideal theories of rationality and meaning developed in epistemology and economics because in human life deliberation takes time and effort, ignorance and inconsistency do not deter action, and learning takes time and slows with time. This paper examines some theories of mental change with an eye to assessing their adequacy for characterizing realistic limits on change and uses a simple kind of reasoning system from artificial intelligence to illustrate how mechanical concepts, including mental inertia, force, work, and constitutional elasticity, provide a new language and formal framework for analyzing and specifying limits on cognitive systems.

1 Vive la Résistance

The ideal actors on the stage of human imagination exhibit courageousness, decisiveness, integrity, generosity, an ability to think clearly and rapidly as they act, and an ability to change direction instantly should danger or opportunity warrant. We admire heroes and heroines and tell their stories partly to celebrate their attainment of these qualities, for many people exhibit courageousness, decisiveness, integrity, generosity, and clear thinking in small matters, and so appreciate the joy they imagine the hero and heroine must feel in larger matters. But an ability to change direction instantly? The size and profitability of the self-help section of almost any bookstore attests to the trouble most people have in changing their

J. Doyle (✉)
Department of Computer Science, North Carolina State University, Raleigh, NC, USA
e-mail: Jon_Doyle@ncsu.edu

B.-O. Küppers et al. (eds.), *Evolution of Semantic Systems*,
DOI 10.1007/978-3-642-34997-3_7, © Springer-Verlag Berlin Heidelberg 2013

behavior. Indeed, some find it easier to change behavior in large ways than in small, but few find change easy, either to effect or to accept.

The prominence of resistance to change in human nature raises doubts about the standard conception of ideal rationality in thought and action. The foundations of the ideal rationality put forward by decision theory and economics suppose that any bit of new information can lead to arbitrarily large changes arbitrarily quickly. One can easily tell someone "Everything you know is wrong!", but few outside of fantastic fiction can act on such advice even if they believe it. Real people can accept new information without connecting it with prior beliefs to draw new consequences or to notice inconsistencies and must search and think carefully to ensure that they do not overlook some relevant fact in making decisions.

Resistance to change therefore impacts the very nature of human belief. Inferences based on mundane prejudices and stereotypes permeate human knowledge and provide commonly useful conclusions unless one makes the effort to distinguish the atypical aspects of the current circumstances. Failure to find or acknowledge these atypical aspects thus leads to systems of belief corrupted by inappropriate assumptions. More generally, humans organize knowledge into numerous hierarchies of abstractions that reflect commonalities and differences in meanings. Failure to find or acknowledge commonalities and to restructure abstractions to reflect them produces conceptual organizations that omit useful connections between concepts and so impede effective recognition of and response to changing circumstances.

A robust theory of rational reasoning and action falls short without a good account of the origins and character of resistance to change. Philosophy and psychology have developed several ways of understanding resistance to change, especially the notions of habit, refraction, and entrenchment, but suffer a striking omission: the long-standing notions of inertia and force in common informal usage in everyday descriptions. Mathematicians from Galileo to Euler showed how to use these concepts precisely in characterizing the behavior of physical bodies, but psychology has suffered from the lack of such concepts except for use in informal metaphor and analogy. One sees this lack clearly in the case of theories of ideal rationality, in which the characteristic unbounded response to bounded impetus shows none of the proportional response embodied in mechanical laws of force and inertia.

The following presents an account of resistance to change in psychological systems that reveals an underlying mechanical character to psychology. This account employs an extension of the mathematical formulation of mechanics to cover minds and persons with mechanical minds and bodies. As detailed in [14], which develops the formalism in axiomatic terms along with the appropriate mathematical structures, the extended mechanics separates the fundamental properties and structures of position, motion, mass, and force from special characteristics of continuous space and extends these notions to hybrid systems that combine different types of positions, masses, and forces. The presentation here omits the axiomatics in favor of an annotated example of mechanics in a specific type of reasoner developed in artificial intelligence.

2 Active and Passive Resistance

We set the stage for an examination of the mechanical account of resistance to change by reviewing briefly the principal accounts of resistance developed in mental philosophy and psychology without reference to mechanical notions, namely the notions of habit and refraction.

Although conscious thought focuses on our deliberate reasoning and action, habits form a major element of both thought and action, to an extent that some have regarded most or all of thinking as occurring through the action of complex sets of habits. Hume famously saw the foundations of reasoning in terms of experiential associations between one condition and another that developed customs or habits of thinking. Behaviorist psychology later expanded on this notion to interpret all or almost all behavior as occurring through the combined operation of sets of habits, with positive or negative reinforcement modifying behavior by modifying mediating habits. Artificial intelligence adapted notions of habits by employing collections of automated antecedent-consequent rules [25] to replicate certain types of reasoning, and by formalizing behavioristic action in terms of hierarchies of routine behavior [2, 38].

Computational rule engineers viewed the problem of training more broadly than the behaviorists, typically looking first for human informants to state concepts, conditions, and rules explicitly, followed by iterative tailoring of behavior by the engineer or by means of automated learning mechanisms. Modern neurophysiology reinforces the importance of habit by identifying neurons as stimulus-response units and by observing how repetitive usage patterns shape the stimulus sensitivity, the response function, and the network of neural connections.

Development of ideas of habit across the years has yielded an appreciation of the power of individually simple and specialized habits to work together to produce complex and sometimes intelligent behavior, to the point that earlier conceptions of human intelligence as formed mainly by deliberate reasoning modulated by peripheral influences of habits have been upended into conceptions of human intelligence as primarily habitual behavior modulated by occasional elements of deliberate reasoning. Although this transforms the notion of habit into something much more complex and subtle than Hume and others might have had in mind, this deeper understanding of the power of habit has not been accompanied by a similar appreciation of the limits such complex habits place on the power of the reasoner to change thought and action.

Of course, people are not mere creatures of habit; they are stubborn too, exhibiting refractory and willful behavior of sorts lamented throughout history. Human refraction involves active resistance to imposed change that seeks to nullify the imposition and, in willful behavior, maintain or even strengthen current attitudes and activities.

Few philosophers have devoted much attention to the nature of active stubbornness or refraction, with Shand [34] a prominent exception in his discussion of reactions generated by various circumstances in persons of different mental character. Active resistance plays a greater role in artificial intelligence, notably in Minsky's [24] theory of the "society of mind," but few have sought to formalize refraction directly. The most relevant formalizations do not address reaction as much as they address conflict, specifically decision making in which mental subagencies argue with each other [4] or have conflicting preferences [15]. Although one can view refraction through these conceptual lenses, this view does not capture the relation between the imposed change and the reaction to it.

3 Formalizing Resistance to Change

Natural philosophers formalized resistance to change in a variety of familiar ways. Physicists identify the notion of inertia as a property of all matter that characterizes resistance to change of motion. Elasticity characterizes forces generated to restore deformed configurations to undeformed ones, such as the force proportional to displacement characterized by Hooke's law for springs. Friction characterizes forces generated by motion that act against the motion. These formalizations of resistance to change have been unavailable to mental philosophy and artificial intelligence, and this unavailability has led these fields to develop formal methods useful for characterizing resistance to change in phenomenological terms. The following briefly recounts these formal methods.

To simplify the discussion, we initially restrict attention to change of belief in logically conceived reasoners with perfect inferential abilities, and later widen our view to aspects of mental states other than belief and to change in nonideal reasoners. Perfect inferential abilities means that the reasoner knows all consequences of its beliefs. In such theories, one formalizes states of belief as deductively closed and consistent sets of beliefs or "theories" $A = Cn(A)$. General changes of mind can thus take one set A into a new set A' that involves adding and removing multiple statements.

A further simplification analyzes complicated changes into smaller changes that begin with changes due to adding or removing a single belief. There are three types of such smaller changes. One can *add* a statement x to A, denoting the result as $A + x$, simply by taking the deductive closure of the set A extended by x, that is, defining

$$A + x = Cn(A \cup \{x\}). \tag{1}$$

Obviously, simple addition is inappropriate if x is inconsistent with A, so the more useful and general types of change attempt to ensure a consistent result. The operation of *contraction*, denoted $A \mathbin{\dot{-}} x$ removes x from A if possible, and the operation of *revision*, denoted $A \mathbin{\dot{+}} x$, consistently adds x to A by removing conflicting beliefs if necessary. The commonly intended connection between contraction, revision, and addition is given by the Levi identity

$$A \mathbin{\dot{+}} x = (A \mathbin{\dot{-}} \neg x) + x, \tag{2}$$

stating that revision by x should be the same as adding x after first removing anything contradicting x.

Simply naming these types of operations does not define them, inasmuch as one might be able to remove a statement from a theory in many different ways, up to and including removal of all statements except the tautologies, which cannot ever be removed from deductively closed theories. Accordingly, the formal development of these operations begins by understanding the types of changes that fit with each of these operations.

We will divide the phenomenological approaches to analyzing mental change into two subclasses, *comparative* theories that provide means comparing the relative size of different changes, and *characteristic* theories that seek to characterize the superficial properties of the starting and ending states connected by changes. We will later contrast these phenomenological approaches with *substantial* theories that seek to obtain comparative and characteristic properties of revision from more detailed assumptions about the substance and structure of mental states.

3.1 Comparative Theories of Mental Change

The guiding intuition of the comparative theories is Quine's [29] principle of *minimum mutilation*, in which one seeks the "smallest" change that accomplishes the purpose of the operation.

A simple and early example of the comparative approach is Rescher's [32] interpretation of hypothetical and counterfactual reasoning, which assumed a weak preference preordering (reflexive and transitive relation) over all sets of beliefs. To determine what conditions hold in some hypothetical or counterfactual situation, one examines maximal subsets of current beliefs consistent with the hypothetical or counterfactual hypothesis (such as possible contractions) and tries to find subsets that are maximally preferred among these. Rescher called these preferred maximal consistent subsets. One then looks to see if combining each preferred maximal consistent subset with the hypothesis of the hypothetical yields the conclusion of the hypothetical.

Lewis [22] based his semantics for counterfactuals on the notion of *comparative similarity* relations among possible worlds. A comparative similarity relation consists of a ternary relation

$$X \preceq_W Y \tag{3}$$

over possible worlds such that each binary relation $\preceq_W$ is a preordering of possible worlds, that is, satisfies the condition

$$X \preceq_W X \tag{4}$$

for each world X and the condition that

$$X \preceq_W Z \tag{5}$$

whenever

$$X \preceq_W Y \preceq_W Z \tag{6}$$

for each X, Y, Z. Such a preorder provides a notion of similarity relative to each possible world if it satisfies the additional requirement that the preorder for each world be origin-minimizing, that is, assigning minimal rank to its "center" or "origin", for formally, that

$$W \preceq_W X \tag{7}$$

for each W and X.

These requirements encompass similarity obtained from a distance function or metric d by defining

$$X \preceq_W Y \tag{8}$$

to hold iff

$$d(W, X) \leq d(W, Y), \tag{9}$$

but the axioms of comparative similarity do not require any structure of distance, much less Euclidean distance, on the set of possible worlds. A comparative similarity relation provides a way to compare changes with respect to the same origin, but need not provide any way of comparing changes with different starting points. One can capture that broader range of comparisons instead by considering a preorder over all pairs of worlds, namely a binary relation $\preceq$ satisfying the conditions

1. $(X, Y) \preceq (X, Y)$,
2. $(X, Y) \preceq (X'', Y'')$ whenever $(X, Y) \preceq (X', Y') \preceq (X'', Y'')$, and
3. $(X, X) \preceq (X, Y)$ (or even the stronger condition $(X, X) \preceq (Y, Z)$).

One easily verifies that each such a preorder over pairs of worlds induces a comparative similarity relation by defining

$$Y \preceq_X Z \tag{10}$$

to hold just in case

$$(X, Y) \preceq (X, Z)\,. \tag{11}$$

Comparative similarity relations, whether based at worlds or over pairs of worlds, all admit numerical representations in the usual way. We say a real-valued binary function s represents a point-based similarity relation $\preceq$ just in case we have that

$$s(X,Y) \leq s(X,Z) \quad \text{whenever} \quad Y \preceq_X Z \ . \tag{12}$$

We say that s represents a pair-based similarity relation $\preceq$ just in case we have that

$$s(X,Y) \leq s(Z,W) \quad \text{whenever} \quad (X,Y) \preceq (Z,W) \ . \tag{13}$$

Obviously such representations are purely ordinal in character and can make comparable alternatives that were incomparable in the represented partial preorder. We may thus regard the formalism of pair-based comparative similarity as a simple (even simple-minded) means with which to describe different levels of difficulty when moving from one world to another.

Although the notion of similarity of worlds does not in itself rest on any notion of difficulty or of resistance to change, at least some notions of similarity and dissimilarity track the ease and difficulty of changing from one mental state to another, as will be seen later.

3.2 Characteristic Theories of Mental Change

To go further than the abstract notion of comparative similarity, one must be more specific about the nature of mental states. One important example is the theory of Alchourrón, Gärdenfors, and Makinson (AGM) [1, 16], who laid down axioms intended to characterize the observable properties of a notion of logically rational belief revision. Their axioms for contraction are as follows.

($\dot{-}$1) $A \dot{-} x$ is a theory whenever A is
($\dot{-}$2) $A \dot{-} x \subseteq A$
($\dot{-}$3) If $x \notin Cn(A)$, then $A \dot{-} x = A$
($\dot{-}$4) If $\nvdash x$, then $x \notin Cn(A \dot{-} x)$
($\dot{-}$5) If $\vdash x \leftrightarrow y$, then $A \dot{-} x = A \dot{-} y$
($\dot{-}$6) $A \subseteq Cn((A \dot{-} x) + x)$ whenever A is a theory
($\dot{-}$7) $(A \dot{-} x) \cap (A \dot{-} y) \subseteq A \dot{-} (x \wedge y)$ whenever A is a theory
($\dot{-}$8) If $x \notin A \dot{-} (x \wedge y)$, then $A \dot{-} (x \wedge y) \subseteq A \dot{-} x$ whenever A is a theory

These axioms mainly state fairly intuitive conditions, for example, that $A \dot{-} x$ is always included in A ($\dot{-}$2); that contraction leaves A unchanged if $x \notin Cn(A)$ ($\dot{-}$3); that the contraction omits x as long as $\nvdash x$ ($\dot{-}$4); that contractions by equivalent statements yield the same result ($\dot{-}$5); and that adding x to the result of contracting by x only yields conclusions present prior to the contraction ($\dot{-}$6).

AGM showed that a range of natural revision functions based on ideas similar to Rescher's preference-based revisions satisfy all their axioms, with different types of AGM revisions corresponding to the way revision proceeds when there are multiple maximal consistent subsets over which Rescher revision preferences operate. In fact, AGM revision implicitly embodies one revision preference also implicit in Rescher's maximize-first approach, namely a uniform preference for believing more over than believing less [10]. This more-is-better preference underlies traditional thought in epistemology, but is better avoided in nonmonotonic reasoning, in which the reasoner bases inference on preferences about whether particular additional beliefs are better than others, or are better than withholding judgment. Both of these theories, then, are reasonably viewed as theories of rational resistance to change.

One can also broaden the AGM conception to apply to revision of beliefs, preferences, and other attitudes [9]. By replacing ordinary logical relations between sentences with the entailment and consistency notions of a Scott information system [33], one can consider AGM-style revisions over any type of mental attitude or entity.

3.3 Collective Versus Component Comparisons

The AGM axioms characterize changes of entire theories, but do not explicitly relate changes affecting different beliefs. Gärdenfors and Makinson [17] studied theory revision from the perspective of individual statements and showed that the outcomes-based AGM approach implies the existence of an ordering of beliefs by relative *epistemic entrenchment*, such that the effect of adopting a new belief is to abandon the least entrenched prior beliefs necessary to ensure consistency. The ordering $x < y$ of two statements in a theory means that one gives up x before y in contractions when one cannot keep both. They formulate the following set of axioms stating that the entrenchment ordering must be transitive ($\leq$1); exhibit a general type of reflexivity ($\leq$2); must give up at least one conjunct when giving up a conjunction ($\leq$3); must regard propositions not in the current beliefs to be minimally entrenched ($\leq$4); and must regard propositions to be maximally entrenched only if they are logically valid ($\leq$5).

($\leq$1) If $x \leq y$ and $y \leq z$, then $x \leq z$
($\leq$2) If $x \vdash y$, then $x \leq y$
($\leq$3) Either $x \leq x \wedge y$ or $y \leq x \wedge y$
($\leq$4) If A is consistent, then $x \leq y$ for all y iff $x \notin A$
($\leq$5) If $x \leq y$ for all x, then $\vdash y$

Gärdenfors and Makinson then prove that the axioms for contraction and entrenchment are equivalent in the sense that every contraction function corresponds to some entrenchment ordering, and every entrenchment ordering corresponds to some contraction function. Formally, $x \leq y$ holds iff either

$$x \notin A \dot{-} (x \wedge y) \quad \text{or} \quad \vdash x \wedge y \ ; \tag{14}$$

correspondingly,

$$y \in A \dot{-} x \tag{15}$$

holds iff

$$y \in A \quad \text{and either} \quad x < x \wedge y \quad \text{or} \quad \vdash x \ . \tag{16}$$

Because of their reflexive, transitive, and complete character, entrenchment orderings can also be viewed as preference orders [10]. In this setting, utility or loss functions that represent these preference orders can also be taken as measures of resistance to changing individual beliefs.

4 Substantial Theories of Mental Change

The Lewis, AGM, and Rescher conceptions provide a framework for formalizing comparisons between changes that let us say that one change is bigger than another. Although one expects to find such comparisons in any approach to understanding habit and refraction, mere comparisons do not provide insight into why one change ranks bigger than another, nor do they provide cardinal measures to quantify just how much bigger one change is than another. To understand why one change is harder than another, and just how much harder it is, we must look beyond mere comparisons to theories that explain difficulty in terms of the substance and organization of mental states. To do this, we again turn to Quine, who painted a picture of substantial origins for mental change in his image of the "web of belief" [30, 31], in which belief revision restructures a network of connections between beliefs, with the resulting changes guided by the principle of minimum mutilation.

One key element of the Quinian conception is that connections between beliefs reflect some type of *coherence* between the beliefs. Coherence in this conception demands at least consistency between beliefs, and in some cases that some beliefs explain others, but in all cases that the overall set of beliefs is the central focus of attention. Yet Quine's picture also carries within it another conception, one in which it is direct connections between beliefs that take center stage. Think of a physical web, whether that of a spider or a fisherman's netting, and one thinks of pokes and tugs moving and stretching smaller or larger portions of the web, and of the elements of the web exhibiting degrees of stiffness, elasticity, and the like. Such an approach to mental change has been explored in artificial intelligence in two ways: *structural* or *foundational* methods based on reasons and dependencies and the foundations they provide for beliefs and other attitudes, and *preferential* methods based on domain-specific preferences among beliefs and how they guide rational reasoning.

4.1 Structural Theories of Mental Change

In foundational organizations for mind, the state of belief exhibits a division of memory into at least two parts, variously identified as axioms and theorems [23], reasons and conclusions [3], implicit belief and explicit belief [21], or constructive belief and manifest belief [8]. The essential commonality in all these conceptions is that some type of base elements generate or provide the foundation for some set of extension elements in the sense that beliefs appear as conclusions only if supported by the base or foundational beliefs.

Truth maintenance systems or, more accurately, reason maintenance systems (RMS) and related dependency-based revision frameworks [3, 35, 36] form the exemplars of foundational psychologies. In such systems, records called *reasons* or *justifications* represent finite traces of past derivations or reasoning steps. Traces of ordinary deductive inferences can be represented by deductive reasons

$$A \ \Vdash \ C \ , \tag{17}$$

read "A gives C", and standard nonmonotonic inferences can be represented by nonmonotonic reasons

$$A \ \setminus\!\!\setminus \ B \ \Vdash \ C \ , \tag{18}$$

read "A without B gives C", informally interpreted as indicating that holding every member of the set A of *antecedent* statements without holding any of the *defeater* statements in B gives each of the *consequence* statements in C. For example, one can represent an inference of the form

"Conclude (c) Sasha can fly
whenever it is believed that (a) Sasha is a bird, and
it is not believed that (b) Sasha cannot fly."

by the reason

$$\{a\} \ \setminus\!\!\setminus \ \{b\} \ \Vdash \ \{c\} \ , \tag{19}$$

and an "axiom" that Sasha is a bird can be represented by the reason

$$\varnothing \ \setminus\!\!\setminus \ \varnothing \ \Vdash \ \{a\} \ . \tag{20}$$

The typical RMS uses these inferential records to construct a set of conclusions that contain all and only the conclusions demanded by the recorded reasons, so that every belief in the set of conclusions is a consequence of some reason for which the antecedents are present (*In*) and the defeaters are absent (*Out*). Formally, a reason

$$A \setminus\!\!\setminus B \Vdash C \tag{21}$$

requires that the set X of extended conclusions satisfies the condition

$$[A \subseteq X \subseteq \overline{B}] \rightarrow [C \subseteq X] , \tag{22}$$

where $\overline{B}$ denotes the set of all statements not in B. We say that

$$A \mathbin{\backslash\!\!\backslash} B \Vdash C \tag{23}$$

is *valid* if

$$A \subseteq X \subseteq \overline{B}. \tag{24}$$

The requirement that each conclusion be supported by some valid reason, which one might call local grounding, represents one position on a spectrum of possible grounding conditions [5, 12]. The original RMS imposed a strong global grounding requirement, namely that each conclusion is supported by a noncircular argument from the base reasons [3]. One can also consider intermediate organizations that divide memory into multiple working sets or locales and require strict grounding of conclusions within each locale but only local grounding across locales [13].

In a foundational approach, one effects changes in conclusions indirectly by making changes in the base beliefs. One adds reasons that generate beliefs rather than simply adding beliefs directly. This requires systematic addition of reasons in the course of reasoning, for example, through comprehensive recording of the reasoning history [19] or using episodic summarization methods such as chunking [20]. The reasons persist in memory until explicitly removed, even as the set of conclusions changes due to changes in the set of base reasons. In fact, by employing reason structures that provide for uniform nonmonotonic defeasibility, one need not ever remove reasons, but instead can defeat any reason marked for removal [4].

In the foundational setting, reasoned contraction involves tracing underlying derivations recursively. To remove some conclusion, one looks to its supporting reason or argument and seeks to defeat some assumption or reason involved in this support, as in the technique of dependency-directed backtracking [3, 35]. The removal of conclusions by support tracing need not be immediate, because when one assumption is removed or defeated, one must reevaluate the reasons by which it supported other statements. If those supported statements have alternative means of support in other reasons, one must repeat the support-tracing process and make further changes until one finally removes the target conclusion.

4.2 Measuring Revision Difficulty

Reason-based revision, as with any concrete computational method, permits analysis of the difficulty of effecting changes in computational terms of numerical measures of the quantities of time and memory required. Such measures have limited appeal as methods of judging difficulty of change, for the wide variability of time and space requirements due to variations in algorithms and architectures means that the difficulty of a particular change has little relation to specific numerical measures of time and space.

Rather than looking to standard computational measures of difficulty, one can instead employ noncomputational comparisons of difficulty of change related more directly to the reason maintenance approach itself [5]. Such comparisons can take the form of Lewis-like relations

$$X \preceq_W Y \tag{25}$$

of comparative similarity of belief states W, X, Y that capture the intent, if not the practice, of reason maintenance revisions. For example, one can compare change difficulty in terms of the set of changed conclusions, so that

$$X \preceq_W Y \quad \text{iff} \quad X \triangle W \subseteq Y \triangle W \ , \tag{26}$$

where $X \triangle Y$ represents the symmetric difference

$$(X \setminus Y) \cup (Y \setminus X) \ . \tag{27}$$

One can also measure change difficulty in terms of the number of changes made, in which case

$$X \preceq_W Y \quad \text{iff} \quad |X \triangle Y| \leq |Y \triangle W| \ . \tag{28}$$

4.3 Exploiting Substantive Measures in Phenomenal Theories

Although phenomenal theories of mental change provided only qualitative comparisons of difficulty of change, one can connect the substantial and phenomenal theories and seek to transfer cardinal measures from substantial to phenomenal theories.

One can interpret psychological systems as reflecting reasons without any implication that the system actually uses representations like reasons in its operation [5,12]. In this approach, one examines the range of states exhibited by the system to identify associations between elements of these states. Reasons correspond to fairly simple conditional and nonmonotonic associations between mental elements. With such interpretations, one can use sets of reasons as a fairly general way of describing the origins of degrees of difficulty of change [7].

More direct connections come from reading entrenchment orderings out of foundational structures and vice versa. Entrenchment orderings arise naturally in foundational psychologies. If one draws a graph in which edges labeled by reasons connect antecedent beliefs with their reasoned conclusions, one can assign levels to the assumptions introduced by nonmonotonic reasons. Assumptions with no antecedents occur at the base, and assumptions with antecedents of a certain level occur at the next level. Following these inferential connections produces an assumption graph that shows which assumptions depend on others. The entrenchment rank of a belief then corresponds roughly to its depth in the assumption graph. Base beliefs are the most entrenched, and to a first approximation beliefs closer to the base beliefs or with multiple derivations are more entrenched than beliefs further from the base or with only one derivation.

One can also go the other way, as Gärdenfors [18] suggests, and read reasons out of entrenchment order. Indeed, del Val [39] proved the equivalence of the coherence and foundations approaches under the assumptions that reasons are beliefs (no nonmonotonic rules), that reasons generate conclusions by deductive closure, that logically equivalent conclusions have identical bases, and that one has an entrenchment ordering on foundational beliefs. This formal equivalence has limited practical import, however, because realistic psychologies do not satisfy these idealizing assumptions. Computational embodiments, for example, rely on finite representations and computations, and neurophysiological embodiments involve large but finite numbers of neurons with relatively small fan-in/fan-out connectivity. Beyond these finiteness considerations, there are practical reasons to regard entrenchment as derived from reasons rather than the reverse [11].

4.4 Preferential Theories of Mental Change

One can connect reasoned foundations with entrenchment in another way as well. As noted earlier, one can interpret nonmonotonic reasons as constraints on sets of beliefs, mandating that certain conclusions be held if certain antecedents are held and others are not. In this interpretation, one can regard reasons as expressing intentions of the reasoner about the content of its mental state, namely that conclusions of valid reasons must be believed.

However, one can also interpret nonmonotonic reasons as expressing preferences about the content of its mental state as well [5,6], somewhat akin to Rescher's revision preferences and the preference-order character of the epistemic entrenchment relation. The preferential content of reasons indicates that indicated nonmonotonic conclusions should be believed in preference to indicated nonmonotonic qualifications, in the sense that sets of conclusions X satisfying

$$[A \not\subseteq X] \vee [A \subseteq X \subseteq \overline{B} \wedge C \subseteq X] \tag{29}$$

are preferred to conclusions satisfying

$$A \subseteq X \not\subseteq \overline{B}\ , \tag{30}$$

which in turn are preferred to conclusions satisfying

$$[A \subseteq X \wedge X \not\subseteq \overline{B} \wedge C \not\subseteq X] \,. \tag{31}$$

The interpretation of reasons as expressing preferences about mental states implies that grounded sets of nonmonotonic conclusions are Pareto optimal, that is, satisfy as large a set of reasoning preferences as possible [6]. It also ties conflicting intuitions about the range of possible revision methods to the problems of group decision making [15].

4.5 Substance and Origins of Change

The substantive theories of mental change go beyond the ideals of pure logical and economic rationality by relating the shape of mental changes to the origins of mental changes. The AGM theory of belief revision, for example, says nothing about how the entrenchment order changes, only that it exists and guides revisions, although some subsequent theories treat some revisions as shifting the base of the entrenchment ranking. Similarly, standard theories of economic rationality say much about how probabilistic expectations change with new evidence, but almost nothing about how or when preferences change. Artificial intelligence addresses these questions with theories of problem solving in which reducing goals to subgoals changes the preferences of the reasoner, in which chunking or related mechanisms change the set of base reasons and hence the foundations of mental states, and in which base and derived reasons express preferences that shape the revisions they induce.

5 A Mechanical Perspective

Even though one can interpret a variety of systems in terms of reasons, reasons still represent only one means for assessing difficulty of change. To find a more general identification of generators of difficulty in change, we look to mechanics and its notions of mass, inertia, and constitutive resistive forces. In the present discussion, the base reasons of the foundations approach constitute the mass of the reasoner, the derived conclusions constitute its position, changes in conclusions constitute its velocity, and changes in base reasons constitute its change in mass. Reasoning in such reasoners generates forces that depend on the character of the reasoner, on how the reasoner conducts volitional and deliberative processes.

Everyday references to habit, refraction, and entrenchment make informal use of mechanical notions in addition to the neutral and psychological language used above. In this long-standing but informal usage, refraction becomes reaction to force, and entrenched habits require force to change. This usage probably predates the development of mechanics as a formal, mathematical subject, but ever since the Newtonian era, application of mechanical concepts to anything but physical systems has involved only metaphor and analogy.

Recently, however, axiomatic work in rational mechanics and formal work in artificial intelligence have laid the groundwork for a change in the theoretical status of psychological application of mechanical concepts. Artificial psychological systems such as RMS satisfy the axioms of mechanics, and thus constitute systems to which mechanical concepts properly apply without recourse to metaphor or analogy [14].

Modern rational mechanics [26, 27, 37] supports applications to psychology by providing a formal theoretical structure that offers several advantages over the older mechanical tradition. Three of these advantages offer special benefit in formalizing psychology in mechanical terms.

First, rational mechanics cleanly separates general mechanical requirements from the properties of special materials. Modern theories of mechanics regard bodies as subject to *general* laws applying to all types of materials, laws that set out the properties of space, time, bodies, and forces, and that relate forces on bodies to the motions caused by the forces.

None of the central mechanical laws say anything about which forces exist, or even that *any* particular forces exist. Such statements instead come in *special* laws of mechanics, such as laws of *dynamogenesis* that characterize the origin of forces. The most general of these set out the laws of inertial forces. Traditional presentations of mechanics give a distinguished role to this force through the Newton–Euler equation

$$f = \dot{p} , \tag{32}$$

in which one takes the quantity f to aggregate all the non-inertial forces acting on the body. The more general law

$$f - \dot{p} = 0 \tag{33}$$

of the balance of forces places the inertial force $(-\dot{p})$ generated by the mass of the body on an equal footing with other forces, and has, as its fundamental principle the balance of all forces acting on the body, the meaning that all these forces sum to zero.

Other special laws characterize the behavior of special types of materials, ordinarily identified in terms of constraints on bodies, configurations, motions, and forces. For example, one obtains rigid body mechanics from the general theory by adding kinematical constraints that fix the relative distances of body parts; elastic body mechanics comes from adding the assumption that body deformations generate elastic forces dependent on the deformation; and the mechanical theory of rubber comes from modifying the general theory of elastic materials with the configuration-dependent forces characteristic of rubber. Mechanics uses the term *constitutive assumptions* to refer to special laws for particular materials, since each such law reflects an assumption about the constitution of the material. Mechanical practice depends critically on these special laws. Some so-called fundamental laws of physics constitute special laws of "elementary" particles and fields; most of these, however, stand largely irrelevant to the more numerous special laws characterizing ordinary materials. Rigorous derivation of most special laws from "fundamental" properties of elementary particles remains well beyond present theoretical capabilities, even if such derivations exist.

Second, rational mechanics separates much of the structure of ordinary physics from assumptions about continuity of the quantities involved. The modern axioms of mechanics state fundamental properties of mechanical notions with reference only to some basic algebraic and geometric properties of the spaces involved. Ordinary numbers exhibit these properties, but modern algebra and geometry show how many of the familiar properties of numbers required for mechanics also occur in discrete and finite structures. The resulting algebraic structures for space, mass, and force have much the same character as standard conceptions, although the broadened mechanics allows the possibility of different types of mass, much as pre-relativistic classical mechanics regarded inertial and gravitational mass as different mechanical properties that mysteriously had proportionate values. This broadening of the notion of mass means that mass and velocity sometimes combine in different ways than in traditional mechanics.

One can thus purge the axioms of mechanics of continuity assumptions in the same way one purges them of constitutive properties of special materials, and so obtain a mechanics covering discrete and continuous systems, either separately or in combination, in which the usual continuity assumptions need hold only for the usual physical subsystems.

Third, rational mechanics provides a formal characterization of the notion of force in a way that covers psychological notions. Mechanics did not provide any axiomatic characterization of force until the middle of the past century, well after Hilbert made formalization of all of physics one of his famous problems in 1900. As formalized by Noll [26], the theory of force takes on a generality that covers both physics and psychology. The general laws of forces, for example, state the additivity, balance, and frame-indifference of forces on each body. These state that the force on a body is the sum of all the forces on its disjoint subbodies (additivity); that all forces acting on a body add to zero (balance); and that the true force on a body does not depend on the observer (frame-indifference).

Obtaining the full benefits of mechanical concepts in psychology will require considerable mathematical work to provide an analysis of discrete and hybrid mechanical systems that matches the theoretical and methodological power of modern continuum mechanics, as well as work to embed mechanical concepts in languages for specifying and analyzing designs for reasoners.

6 Effort in Discrete Cognition

To illustrate the mechanical perspective on mind, we sketch a highly simplified illustration based on RMS. In this illustration we regard cognitive states as including sets of discrete beliefs, preferences, desires, and intentions, as well as reasoning rules or habits, especially rules of a sort we will regard as reasons or justifications. We write $\mathcal{D}$ to denote the set of all possible attitudes making up cognitive states, so that $\mathcal{P}(\mathcal{D}) = 2^{\mathcal{D}}$ represents all possible states $S \subseteq \mathcal{D}$.

6.1 Kinematics and Dynamics

We represent cognitive states using vectors in the binary vector space $\mathbb{D} = (\mathbb{Z}_2)^{\mathcal{D}}$ over scalars $\mathbb{Z}_2$. In the RMS terminology, 1 means *In* and 0 means *Out*. We define

$$\mathbf{0} = \varnothing = (0, 0, \ldots), \quad \mathbf{1} = \mathcal{D} = (1, 1, \ldots) \ , \tag{34}$$

and

$$\mathbf{1} - x = \overline{x} = \mathcal{D} \setminus x \ . \tag{35}$$

Addition corresponds to symmetric difference, so

$$x + x = \mathbf{0} \tag{36}$$

and

$$x - y = x + y \ . \tag{37}$$

Pointwise multiplication corresponds to intersection, so that

$$xy = x \cap y \ . \tag{38}$$

For infinite $\mathcal{D}$, we consider only vectors representing the finite and cofinite subsets. In this space, orthogonal transformations are permutations.

We write $x_t \in \mathbb{D}$ to denote the position at discrete instant t and

$$\dot{x}_t = x_t - x_{t-1} \tag{39}$$

to denote the trailing velocity. This trailing velocity corresponds to the change signals used in some automated reasoners in triggering rules [3]. The leading acceleration

$$\ddot{x}_t = \dot{x}_{t+1} - \dot{x}_t \tag{40}$$

$$\begin{tikzcd}
\ddot{x}_{t-1} \arrow[dr] & & \boxed{\ddot{x}_t} \arrow[dr] & & \ddot{x}_{t+1} \\
& \boxed{\dot{x}_t} \arrow[ur] \arrow[dr] & & \dot{x}_{t+1} \arrow[ur] \arrow[dr] & \\
x_{t-1} \arrow[ur] & & \boxed{x_t} \arrow[ur] & & x_{t+1}
\end{tikzcd}$$

Fig. 1 The kinematical relationships among position variables in time. The *boxed* quantities denote conventional labels for the quantities of interest at instant t, with a reasoning agent observing x_t and $\dot{x}_t$ and choosing $\ddot{x}_t$

reflects the additions and removals indicated by steps of reasoning. We depict these kinematical quantities in Fig. 1.

We denote the mass at an instant by $m_t \in \mathbb{D}$, and the leading mass flux

$$\dot{m}_t = m_{t+1} - m_t \ . \tag{41}$$

The momentum

$$p_t \in \mathbb{D} \times \mathbb{D} \tag{42}$$

decomposes into separate mass and velocity components as $p_t = (m_t, \dot{x}_t)$. We write the leading momentum change as

$$\dot{p}_t = p_{t+1} - p_t = (\dot{m}_t, \ddot{x}_t) \ . \tag{43}$$

We denote the force at an instant by

$$f_t \in \mathbb{D} \times \mathbb{D} \ . \tag{44}$$

Euler's law of linear momentum then takes the familiar form

$$f_t = \dot{p}_t = (\dot{m}_t, \ddot{x}_t) \ . \tag{45}$$

The total force

$$f_t = f_t^a + f_t^s \tag{46}$$

on the reasoner combines the applied force f_t^a of environment on the reasoner with the self-force f_t^s of the reasoner on itself.

In this setting, it is natural to interpret reasons as conditional invariants of the motion. In particular, an *interval reason*

$$r = A_r \setminus\!\!\setminus B_r \Vdash C_r \setminus\!\!\setminus D_r \ , \tag{47}$$

read "A without B gives C without D", or *antecedents* without *defeaters* gives *consequences* without *denials*, adds its consequences and removes its denials if the reason is *valid* (antecedents *In* and defeaters *Out*).

We obtain the forces generated by such reasons as follows. In the indirect change setting of the RMS, changes to position start with the application of a *mass* force $f_t^a = (\dot{m}_t, \mathbf{0})$ that adds or removes base reasons. To update the position according to a new base reason r, the RMS then generates a *spatial* self-force

$$f_r^s(x_t, m_t, \dot{x}_t) = (\mathbf{0}, \ddot{x}_t) \; . \tag{48}$$

The application of r produces the new position

$$x_{t+1} = \begin{cases} x_t + C_r\overline{x}_t + D_r x_t & \text{if} \quad A_r\overline{x}_t + B_r x_t = \mathbf{0} \\ x_t & \text{otherwise} \end{cases} \tag{49}$$

from which we obtain the velocity

$$\dot{x}_{t+1} = \begin{cases} C_r\overline{x}_t + D_r x_t & \text{if} \quad A_r\overline{x}_t + B_r x_t = \mathbf{0} \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{50}$$

and acceleration

$$\ddot{x}_t = \begin{cases} C_r\overline{x}_t + D_r x_t - \dot{x}_t & \text{if} \quad A_r\overline{x}_t + B_r x_t = \mathbf{0} \\ \dot{x}_t & \text{otherwise} \end{cases} \tag{51}$$

6.2 *Power and Work*

With reasoning mediated by reasons in this way, we can calculate the work and effort expended in reasoning as follows. The *power*

$$P_t = \dot{x}_{t+1} \cdot f_t \tag{52}$$

exerted across interval $(t, t+1)$ is found in the inner product of the force acting across that interval with the velocity across that same interval. The differing temporal subscripts of velocity and force in this formula reflect the difference between leading forces and trailing velocities. We calculate the instantaneous power P_t to be

$$\begin{aligned} P_t &= \dot{x}_{t+1} \cdot f_t && (53) \\ &= \dot{x}_{t+1} \cdot (\dot{m}_t, \ddot{x}_t) \\ &= |(\dot{m}_t, \dot{x}_{t+1}\ddot{x}_t)| && (54) \end{aligned}$$

$$
\begin{aligned}
&= |(\dot{m}_t, \dot{x}_{t+1} - \dot{x}_{t+1}\dot{x}_t)| \\
&= |(\dot{m}_t, \dot{x}_{t+1})| - |(0, \dot{x}_{t+1}\dot{x}_t)|, \qquad (55)
\end{aligned}
$$

here using the norm that counts the number of 1s in a vector corresponding to the ordinary inner product of binary vectors. The work expended across some interval, therefore, is the integral (sum) of the power over that interval, and we may use this mechanical work as a measure of the mental effort of reasoning.

We examine this formula in greater detail to understand how this measure of effort works in the context of reasoning. Different reasoners operate in different ways, and we find that the overall effort of reasoning varies accordingly.

One can divide reasoning models into two types that differ in the treatment of time. In what one might call the *internal time* model, one identifies instants with steps of reasoning, no matter how long separates these steps in the world at large. In what one might call an *external time* model, one regards steps of reasoning as separated by intervals during which the reasoner does nothing. Both of these models offer useful insights.

We consider the internal time model first. This model corresponds best to a notion of deliberate reasoning, in which every step of reasoning involves some change to memory or outlook. In (55) we see that the power expended across a basic unit of time is the change of mass and position minus a cross term $\dot{x}_{t+1}\dot{x}_t$ involving velocity at successive intervals. This cross-term vanishes in deliberate reasoning because normally one does not immediately retract a conclusion one has just drawn, or draw a conclusion one has just retracted; there would be no point to it. In this setting, therefore, we obtain the magnitude of the power expended by the step of reasoning by

$$
\begin{aligned}
P_t &= |(\dot{m}_t, \dot{x}_{t+1})| \qquad (56) \\
&= |\dot{m}_t| + |\dot{x}_{t+1}|. \qquad (57)
\end{aligned}
$$

In this case, the work of a step of reasoning just adds together the number of changes made in memory and attitudes, so the effort involved in a chain of reasoning steps consists of the cumulative number of changes made in memory and attitudes across the span of reasoning.

In the external time model, we regard steps of reasoning as exerting impulse forces on the reasoner, with the reasoner exhibiting inertial (force-free) motion between steps of reasoning. The "Simple Impulse" table in Fig. 2 illustrates the application of a simple impulse spatial force akin to the internal-time model just discussed. This impulse expends an effort of $|\alpha|$ in the time step in which it is applied, according to the preceding calculation. In the subsequent inertial motion, of course, the force f_t vanishes, and so by (53) the power vanishes as well, so the total effort expended in a chain of reasoning steps again equals the cumulative sum of the number of changes to memory and attitudes, with the inertial motions doing no work.

Fig. 2 Kinematic quantities $(x, \dot{x}, \ddot{x})$, power (p), and total effort in two forms of spatial impulse motion starting from rest at the origin location

Time step	Simple Impulse				Up and Down			
t	x	$\dot{x}$	$\ddot{x}$	P	x	$\dot{x}$	$\ddot{x}$	P
0	$\mathbf{0}$	$\mathbf{0}$	α	$\lvert\alpha\rvert$	$\mathbf{0}$	$\mathbf{0}$	α	$\lvert\alpha\rvert$
1	α	α	$\mathbf{0}$	0	α	α	α	0
2	$\mathbf{0}$	α	$\mathbf{0}$	0	α	$\mathbf{0}$	$\mathbf{0}$	0
3	α	α	$\mathbf{0}$	0	α	$\mathbf{0}$	$\mathbf{0}$	0
4	$\mathbf{0}$	α			α	$\mathbf{0}$		
Work	$\lvert\alpha\rvert$				$\lvert\alpha\rvert$			

Inertial motion takes a cyclic form in the discrete space $\mathbb{D}$ due to the algebraic characteristic that $x + x = \mathbf{0}$. As Fig. 2 indicates, inertial motion with velocity α starting from a position α thus traverses the trajectory $\alpha, \mathbf{0}, \alpha, \mathbf{0}, \ldots$. It is certainly not commonplace to think of reasoners as cycling the last set of conclusions in this way. In standard artificial intelligence mechanizations, one instead regards step of reasoning as changing the set of conclusions from one set to another and then leaving it there until the next step of reasoning, as in the internal-time picture of motion. Accommodating this expectation requires one to modify the simplistic picture of reasoning seen in the internal time model.

One easily obtains a more familiar picture of reasoning by regarding steps of reasoning as exerting two impulses, corresponding to the rising and falling edges of a pulse, as depicted in the "Up and Down" table of Fig. 2. That is, the force of the first half of a step of reasoning changes the velocity so as to effect the desired change of position, and the force of the second half of the step of reasoning changes the velocity back to zero by simply reversing (repeating) the thrust of the first half. This produces a pattern of motion of start–stop steps separated by zero-velocity intervals. This start–stop pattern of forces is in fact the pattern of reason forces, in which the frictional force component—$\dot{x}$ provides the falling impulse. This does not involve twice the mechanical effort of the internal time and simple external time pictures, however, because the falling impulses, matched with zero velocities in (54), contribute nothing to the cumulative effort.

Note that mechanical work only measures the effort of making the change itself and does not include any effort involved in evaluating the applicability of some reasoning habit, of searching for the appropriate inference to perform, or of deciding among alternative inferences, if these activities are not effected by means of reasoning steps themselves. If generating a force α requires effort $|\alpha|$, for instance, then the Simple Impulse motion involves an additional effort of $|\alpha|$, while the Up and Down motion involves an additional effort of $2|\alpha|$. The effort associated with such activities, however, depends on the organization and realization of the mind. For example, there need be no effort involved in producing the falling impulse of Up and Down as this value is already available as the velocity. Or for another example, evaluating the applicability of a set of reasons by a serial scan of databases of reasons and conclusions likely involves more effort than by parallel evaluations conducted by reasons wired together like neurons or Boolean circuits.

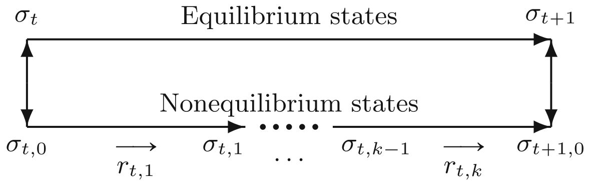

Fig. 3 Reasoned decomposition of a transition between equilibrium states σ_t and σ_{t+1} into a series of "microtransitions" between microstates $\sigma_{t,i}$ effected by a series of reason applications $r_{t,i}$

6.3 Refraction and Elasticity

The effort of reasoning also involves the work of focusing attention on reaching the intended conclusions and on not pursuing distractions arising in the course of reasoning. Autonomous habits of reasoning or perceptions can exert forces of distraction on the reasoner and forces that move it either back toward where it started or merely off the intended path. The reasoner must notice these backwards steps and diversions and force itself back onto the intended path. One pays a double price for refraction and distraction, as both these and the refocusing actions do work and add to the total effort of reasoning. The same holds true for switching between separate reasoning activities being pursued simultaneously, for the effort of switching contexts adds to the efforts attributable to the separate reasoning activities.

The effort involved in a single step of reasoning also enters into the comparisons made by coherence-based conceptions of mental change indirectly because these large-scale changes to mental state typically involve an extended process of smaller reasoning steps. Mechanically, the coherence-based changes take the form of elastic response. The stipulated change, or step of reasoning that triggers a change to a new coherent state of mind, consists of an external deformation. This deformation requires a change to a new "relaxed" or equilibrium state. In RMS revision, this relaxation process consists of a sequence of smaller reasoning steps, which we depict in Fig. 3. As the base reasons constituting the mass of the reasoner grows through learning and experience, the time needed to effect revisions can also grow as revisions involve more and more reasons.

Nonmonotonic reasoning exhibits an elastic character even more strongly. One can view the defeat of an otherwise valid nonmonotonic reason as producing a deformed configuration of the reasoner. Defeat or removal of this defeater forces restoration of the state of mind prior to imposition of the deformation, with the nonmonotonic reason acting like a spring element in shaping the mental configuration. The character of such elastic response is more complex in reasoning than in simple physical materials like springs in that RMS-based reasoners search for equilibria in both relaxation and restoration processes, and the equilibrium resulting from a restoration need not coincide with the one existing prior to the deformation. One must look at more complicated physical systems than individual springs to see similar indeterminacy of equilibrium transitions.

7 Conclusion

Modern rational mechanics provides concepts of mass, force, work, and elasticity that constitute a new analytical framework and vocabulary for characterizing limits to mental change. These concepts go beyond those available in theories of ideal logical and economic rationality that focus on mere comparative difficulty, and beyond computational and neurophysiological theories that focus on measures only loosely related to mental effort. Mechanics also provides other means for characterizing limitations on mental change, including bounds on the forces exerted in reasoning and forms of constitutional rigidity (see [14]).

References

1. Alchourrón, C., Gärdenfors, P., Makinson, D.: On the logic of theory change: partial meet contraction functions and their associated revision functions. J. Symbolic Logic **50**, 510–530 (1985)
2. Brooks, R.A.: A robust layered control system for a mobile robot. IEEE J. Robot. Automat. **2**(1), 14–23 (1986)
3. Doyle, J.: A truth maintenance system. Artif. Intell. **12**(2), 231–272 (1979)
4. Doyle, J.: A model for deliberation, action, and introspection. AI-TR 581, Massachusetts Institute of Technology, Artificial Intelligence Laboratory (1980)
5. Doyle, J.: Some theories of reasoned assumptions: an essay in rational psychology. Technical Report 83–125, Department of Computer Science, Carnegie Mellon University, Pittsburgh (1983)
6. Doyle, J.: Reasoned assumptions and Pareto optimality. In: Proceedings of the 9th International Joint Conference on Artificial Intelligence, pp. 87–90. Morgan Kaufmann, San Francisco (1985)
7. Doyle, J.: Artificial intelligence and rational self-government. Technical Report CS-88-124, Carnegie-Mellon University Computer Science Department (1988)
8. Doyle, J.: Constructive belief and rational representation. Comput. Intell. **5**(1), 1–11 (1989)
9. Doyle, J.: Rational belief revision. Presented at the third international workshop on nonmonotonic reasoning, Stanford Sierra Camp, CA, June 1990
10. Doyle, J.: Rational belief revision (preliminary report). In: Fikes, R.E., Sandewall, E. (eds.) Proceedings of the Second Conference on Principles of Knowledge Representation and Reasoning, pp. 163–174. Morgan Kaufmann, San Mateo (1991)
11. Doyle, J.: Reason maintenance and belief revision: foundations vs. coherence theories. In: Gärdenfors, P. (ed.) Belief Revision, pp. 29–51. Cambridge University Press, Cambridge (1992)
12. Doyle, J.: Reasoned assumptions and rational psychology. Fundamenta Informaticae **20**(1–3), 35–73 (1994)
13. Doyle, J.: Final report on rational distributed reason maintenance for planning and replanning of large-scale activities (1991–1994). Technical Report TR-97-40, ADA328535, Air Force Research Laboratory (1997)
14. Doyle, J.: Extending Mechanics to Minds: The Mechanical Foundations of Psychology and Economics. Cambridge University Press, London (2006)
15. Doyle, J., Wellman, M.P.: Impediments to universal preference-based default theories. Artif. Intell. **49**(1–3), 97–128 (1991)

16. Gärdenfors, P.: Knowledge in Flux: Modeling the Dynamics of Epistemic States. MIT, Cambridge (1988)
17. Gärdenfors, P., Makinson, D.: Revisions of knowledge systems using epistemic entrenchment. In: Vardi, M.Y. (ed.) Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge, pp. 83–95. Morgan Kaufmann, Los Altos (1988)
18. Gärdenfors, P.: The dynamics of belief systems: foundations vs. coherence theories. Revue Internationale de Philosophie **172**, 24–46 (1990)
19. de Kleer, J., Doyle, J., Steele Jr., G.L., Sussman, G.J.: AMORD: explicit control of reasoning. In: Proceedings of the 1977 Symposium on Artificial Intelligence and Programming Languages, pp. 116–125. ACM, New York (1977)
20. Laird, J.E., Rosenbloom, P.S., Newell, A.: Chunking in soar: the anatomy of a general learning mechanism. Mach. Learn. **1**(1), 11–46 (1986)
21. Levesque, H.J.: A logic of implicit and explicit belief. In: Proceedings of the National Conference on Artificial Intelligence, American Association for Artificial Intelligence, pp. 198–202. AAAI Press, Palo Alto, CA, USA (1984)
22. Lewis, D.: Counterfactuals. Blackwell, Oxford (1973)
23. McCarthy, J.: Programs with common sense. In: Proceedings of the Symposium on Mechanisation of Thought Processes, vol. 1, pp. 77–84. Her Majesty's Stationery Office, London (1958)
24. Minsky, M.: The Society of Mind. Simon and Schuster, New York (1986)
25. Newell, A., Simon, H.A.: Human Problem Solving. Prentice-Hall, Englewood Cliffs (1972)
26. Noll, W.: The foundations of classical mechanics in the light of recent advances in continuum mechanics. In: Henkin, L., Suppes, P., Tarski, A. (eds.) The Axiomatic Method, with Special Reference to Geometry and Physics; Proceedings of an International Symposium held at the University of California, Berkeley, December 26, 1957–January 4, 1958. Studies in Logic and the Foundations of Mathematics, pp. 266–281. North-Holland, Amsterdam (1958). Reprinted in [28]
27. Noll, W.: Lectures on the foundations of continuum mechanics and thermodynamics. Arch. Rational Mech. Anal. **52**, 62–92 (1973). Reprinted in [28]
28. Noll, W.: The Foundations of Mechanics and Thermodynamics: Selected Papers. Springer, Berlin (1974)
29. Quine, W.V.: Philosophy of Logic. Prentice-Hall, Englewood Cliffs (1970)
30. Quine, W.V.O.: Two dogmas of empiricism. In: From a Logical Point of View: Logico-Philosophical Essays, 2nd edn., pp. 20–46. Harper and Row, New York (1953)
31. Quine, W.V., Ullian, J.S.: The Web of Belief, 2nd edn. Random House, New York (1978)
32. Rescher, N.: Hypothetical Reasoning. North Holland, Amsterdam (1964)
33. Scott, D.S.: Domains for denotational semantics. In: Nielsen, M., Schmidt, E.M. (eds.) Automata, Languages, and Programming: Ninth Colloquium. Lecture Notes in Computer Science, vol. 140, pp. 577–613. Springer, Berlin (1982)
34. Shand, A.F.: The Foundations of Character: Being a Study of the Tendencies of the Emotions and Sentiments, 2nd edn. Macmillan and Company, London (1920)
35. Stallman, R.M., Sussman, G.J.: Forward reasoning and dependency-directed backtracking in a system for computer-aided circuit analysis. Artif. Intell. **9**(2), 135–196 (1977)
36. Sussman, G.J., Stallman, R.M.: Heuristic techniques in computer-aided circuit analysis. IEEE Trans. Circ. Syst. **CAS-22**(11) (1975)
37. Truesdell, C.: A First Course in Rational Continuum Mechanics, vol. 1. Academic, New York (1977)
38. Ullman, S.: Visual routines. AI Memo 723, MIT AI Lab, Cambridge (1983)
39. del Val, A.: Non-monotonic reasoning and belief revision: syntactic, semantic, foundational, and coherence approaches. J. Appl. Non-Classic. Logics **7**(2), 213–240 (1997). Special Issue on Inconsistency-Handling