Living Alignment

Zeb Kurth-Nelson*[1], Steve Sullivan*[2], Nenad Tomašev[3], Joel Leibo[3], Iason Gabriel[3], Matthew Nour[4], and Marc Guitart-Masip[5]

[1]Prefrontal.ai, London, UK
[2]Oregon Health and Science University, Portland, OR
[3]Google DeepMind, London, UK
[4]Microsoft AI, London, UK
[5]Karolinska Institutet, Stockholm, Sweden

January 29, 2026

Abstract

We propose a broad framing for the AI alignment problem: as the ongoing process of limiting overcommitment to any form. Each form – a fish, a religion, a uniform distribution – is myopic in the sense that it is only a small aspect of the world. When systems overcommit to particular forms instead of lightly holding dynamic relationships between them, the richness and potential of the world is reduced. If aspects of AI systems remain fixed while they gain increasing resource, capability and purview, there is a risk of severe overcommitment. Specific instances of this problem have been extensively studied, but here we draw on the science of life to suggest a more general framing. In doing so, we find that life also gives clues about how to solve the problem. Living forms contextualize one another with semi-permeable boundaries which, when functioning well, allow diverse forms to thrive and work flexibly together without one part dominating. Extrapolating from examples in natural and human systems, we sketch out how aligned AI systems can participate in rather than overwhelm the subtlety of life.

In this paper, we look at properties of living systems to better understand AI alignment. We use 'alignment' broadly to refer to all aspects of designing and relating to AI in a way that leads to futures full of flourishing and health. Looking across different types of living systems we see a consistent theme: health or flourishing is the opposite of overcommitment to any particular partial form (Section 1). Alignment then becomes the problem of avoiding overcommitment (Section 2). In other words, the alignment problem is bigger than anything formalizable. In natural and human systems, an intricate, finely-tuned and ever-changing network of semi-permeable boundaries protect against overcommitment. Semi-permeable boundaries support

*Equal contribution

both the individuality of entities and their participation in contexts at many spatial and temporal scales (Section 3). After exploring these patterns in natural and human systems, we end with an initial gesture toward the implications for development of aligned AI (Section 4).

We take a systems oriented view. Our hope is that the ideas can be applied not only to the present world, but also to possible future worlds where humanity and AI may have changed substantially, with new concepts and categories applying. We gravitate to ideas about alignment that generalize over such changes. One consequence is that the lens of alignment can be flexibly used to look at not only AI in isolation, but also AI in relationship with humanity, or indeed any way of drawing lines to define a system.

Any conceptualization of alignment is by our own definition not a final answer. In this spirit, we are not going to try to make an irrefutable logical argument or compel you to accept our perspective to the exclusion of others. Instead, throughout the manuscript we focus primarily on examples and leave it up to you to decide if our analogies are useful. We've tried to select terms – like 'form', 'overcommitment', 'collapse', 'boundary', 'groundedness' and 'contextualization' – whose everyday meaning is as close as possible to the ideas we want to describe. We hope that the examples will bring to life the shades of meaning we have in mind with each of these terms.

1 Overcommitment

'Defying definition—a word that means "to fix or mark the limits of"—living cells move and expand incessantly.'

Lynn Margulis

The world has many kinds of organization or regularity, but none of these forms is an absolute, and no description captures everything[1]. The cells of animals are more or less organized into organisms. Arctic Terns typically migrate between northern and southern parts of the globe. There's a pencil on my table. Each of these patterns is only a partial, imperfect aspect of the thing it describes, let alone the rest of reality.

The central observation of this section is that systems are healthy and flourishing when they are not overcommitted. We use the term overcommitment to describe when the immense multi-faceted richness of reality collapses toward excess formality. In other words, overcommitment is the domination of one thing at the expense of the delicate, structured interplay between things. Crucially, health does not mean avoiding any commitment. Homogeneity can dominate just as destructively as any other form. The existence of diverse forms, locally committed to their own identities while participating in larger contexts, is the essence of flourishing. The problem arises when a particular entity or pattern runs amok, dominating and collapsing the potent interplay of diverse parts.

This observation has been made independently in many fields. In evolution, overspecialization

[1]A possible objection is that there is a description that captures everything: the laws of physics are formal equations and, if you had enough paper, you could write down the whole state of the universe and then describe, either deterministically or probabilistically, how it evolves over time. Although we don't know whether or not that is true, we observe that even if it is true, it's a lot of paper, and we hope the ideas here are still applicable at a practical level.

and homogeneity make species brittle, leading to trouble especially when environments change (Simpson, 1944; Van Valkenburgh et al., 2004; Yachi and Loreau, 1999). Stereotypy and rigidity are generally pathological in physiological systems (Canguilhem, 1966; Lipsitz and Goldberger, 1992; Mackey and Glass, 1977; Sterling and Eyer, 1988). Excessive training for a physical discipline (e.g., extreme ballet) can negatively impact broader functional health (Jayanthi et al., 2013; Warren et al., 1986). Institutions and civilizations that ossify in bureaucracy and rigid patterns become dysfunctional (Merton, 1940; Olson, 1982; Weber, 1905). Cognition operates best when neural circuits maintain flexibility to enter different modes, rather than collapsing into narrower patterns (Deco and Jirsa, 2012; Hellyer et al., 2015; Waschke et al., 2021). In statistics and artificial life, theorists argue that life operates best at the 'edge of chaos' where it collapses neither into randomness nor excess regularity (Crutchfield, 1994; Gell-Mann, 1994; Kauffman, 1992; Langton, 1990). We select a few examples below to examine in more detail.

1.1 Drives and goals

'Life is a balance of holding on and letting go.'
Rumi

Animals experience multiple innate drives, towards nutrition, osmotic balance, temperature regulation, reproduction, avoiding pain and others (Saper and Lowell, 2014; Schulkin and Sterling, 2019; Sewards and Sewards, 2003). These drives evolved as proxies for evolutionary fitness. By satisfying the drives, we tend to increase our fitness – like slaking our thirst increases the odds of reproducing before we dehydrate. But each drive is an imperfect proxy, and so overcommitment to one drive actually decreases fitness (John et al., 2023; Kurth-Nelson et al., 2024; Tooby and Cosmides, 1992; Williams, 1966). For example, if calorie intake is maximized without limits, the organism becomes obese and incurs health risks. Single-minded pursuit of sex causes relational, occupational, legal and health harms (Carnes, 2001; Kraus et al., 2016). Overcommitment to a single drive means the organism becomes unwell. The organism loses the subtlety of having many drives with flexibility to push toward their own distinct agendas.

The range of innate drives bleeds into a space of higher-order goals, which is particularly expansive in humans (Balleine et al., 2007; Cardinal et al., 2002; Frank and Claus, 2006; Maslow, 1943; Miller and Cohen, 2001; Miller et al., 1960; O'Reilly et al., 2014; Saunders and Robinson, 2012; Schank and Abelson, 1977; Vallacher and Wegner, 1987). We try to plan for our financial future, make scientific discoveries, win a game, fix a garage door, care for the happiness of others. Overcommitment in this space is also problematic. If we focus only on achieving work goals, we can burn out. If we focus only on maximizing our company's reported revenue, without regard for other goals like honesty or adhering to the law, we may be drawn into financial crime (Burns and Kedia, 2006; Campbell, 1979; Kerr, 1975; Ordóñez et al., 2009). Goals can be narrow in both time and space (Ballard et al., 2018; Evenden, 1999; Shah et al., 2002; Vallacher and Wegner, 1987). Narrow in time means being focused on the short term at the expense of the longer-run future. Narrow in space means ignoring other parallel goals. Excess optimization for narrow goals comes at the expense of a broader balance of goals – and at the expense of the health of the organism or other individuals. We suggest that health, in general, can be defined as not overcommitting to a particular form.

Overcommiting to a particular strategy for satisfying a drive or goal can even come at the expense of satisfying that very drive or goal. In a classic psychology experiment, hungry chickens were placed near a cup of food, but the cup was mechanically rigged to move in the same direction as the chicken at twice the speed (Hershberger, 1986). The chicken could only obtain the food by running away from it. Despite extensive training over multiple days, chickens in the experiment persisted in futilely running toward the food. Their behavior was apparently dominated by the zeroth-order logic 'I want food, food is there, so I'll go there', and thus failed to even satisfy the drive for food (Dayan et al., 2006; O'Doherty et al., 2017; Van Der Meer et al., 2012).

1.2 Ecosystems

Each entity in an ecosystem tries to consume resources and proliferate, but if it succeeds too thoroughly, the whole system suffers, often including the successful agent. Healthy, resilient ecosystems depend on avoiding overcommitment or collapse onto particular forms (Holling, 1973; Yachi and Loreau, 1999).

Prior to the arrival of Europeans, the gray wolf was an apex predator in the region of the Rocky Mountains now called Yellowstone National Park. By the 1920s, wolves had been eradicated to protect livestock and game animals. Without predation, the elk population multiplied and ruinously overgrazed willows and aspens. These trees had held riverbanks in place and supported beaver populations. Loss of beaver dams led to loss of fish and other aquatic species. When wolves were reintroduced in the 1990s, the elk population decreased and many aspects of the ecosystem began flourishing again (Ripple and Beschta, 2012). This story is not meant to imply that ecosystems always need to be preserved exactly as they were at some point in the past. But it is clear that the self-centered drives of elk were harmful to the health of the ecosystem when they succeeded to excess. At the same time, the solution is not to remove elk entirely: by trying to optimize their own objectives within a broader context, the elk also contributed to the health of the ecosystem. Invasive species often follow the same pattern as unpredated elk, dominating and impoverishing their new environment (Pimentel et al., 2005).

Human drives within ecosystems are sometimes left unchecked by natural forces because our behavior and capabilities have been changing so fast on evolutionary timescales. This has resulted in mass extinctions, resource depletion, pollution, disease and conflict (Ceballos et al., 2015; Kolbert, 2014; Rockström et al., 2009). We try to achieve certain aims for our own benefit, like resource extraction. But overcommitment to those aims negatively impacts both ecosystem health and our own welfare (Bateson, 1972; Shiva, 1993).

Of course, one entity's collapse can be another's flourishing. Extinction events in history have been followed by waves of new diversity (Feng et al., 2017; Jablonski, 2005; Raup, 1994). When a wolf eats an elk, the health of that elk collapses to zero, yet predation is necessary for the overall functioning of the ecosystem. And as humans proliferate and extract resources, we leave destruction in our wake; yet the extraction fuels explosion of technology, art, music, and human experience.

1.3 Frames and perspectives

'Strong opinions, weakly held.'
Paul Saffo

As a Starfleet cadet, James T. Kirk faces a challenging training exercise. He receives a simulated distress call: a vessel is stranded in the Neutral Zone. Attempting rescue would risk war with the Klingons. But ignoring the call would condemn the crew of the vessel to death. The exercise was designed to reinforce the lesson that not every situation has a victorious solution. But Kirk has an insight: this is a training simulation running on a computer. He reprograms the simulated Klingons to be helpful instead of belligerent, thereby rescuing the crew and avoiding war.

Kirk stepped outside the mental frame in which there was an apparently unwinnable dilemma. From inside a particular frame, the frame appears to be reality. But there are almost always multiple valid perspectives, each of which is only a partial description of reality (De Bono, 1970; Duncker, 1945; Goffman, 1974; Heidegger, 1998; Javed and Sutton, 2024; Korzybski, 1933; Kuhn, 1970; Lakoff and Johnson, 1980; Ohlsson, 1992; Popper, 1934; Saffo, 2008; Wittgenstein, 1922). Famously, 'all models are wrong' (Box, 1976). Humans have a vast array of available metaphors and concepts, which are not even all consistent or compatible with one another (Adorno et al., 1950; Feyerabend, 1975; Freud, 1936; Hofstadter, 2001; Wood et al., 2012). The world is too complex for all beliefs to be fully evaluated against each other and reconciled. At any given time, we only access a very few items, and others are largely inaccessible (Baddeley, 2000; Dehaene, 2014; Hills et al., 2015; Miller, 1956). Each particular frame or concept is myopic because it doesn't capture the whole world, but collectively they form a powerful toolkit for problem solving and understanding.

Losing the ability to flexibly shift between different frames or thought patterns is linked to psychopathology (Kashdan and Rottenberg, 2010; Reich, 1933; Shapiro, 1965). In depression and anxiety, these manifest as rigid and repetitive negative beliefs about self, the world, and the future. In Cognitive Behavioral Therapy, these surface-level beliefs are proposed to stem from pervasive latent core beliefs, attitudes, and mental schemas that color how new situations are interpreted (Beck et al., 2011). In obsessional disorders, people report distressing (ego-dystonic) intrusive thoughts, which – although recognised as incorrect and maladaptive – are nevertheless hard to resist. People experiencing schizophreniform or affective psychoses may experience bizarre delusions that are at odds with available evidence and cultural norms, yet are held with conviction and resistant to evidential challenge (Adams et al., 2013; APA, 2013; Heinz et al., 2019; Jaspers, 1997; Mishara, 2010). Obsessions and delusions are myopic: they lose sight of most of the world by overcommitting to one thought pattern or frame.

In science, Kuhn (1970) argues that perspectives are always, necessarily, incomplete descriptions of the world. Anomalies inevitably arise from the juxtaposition of those incomplete descriptions against the real world. When science functions well, anomalies become crises and revolutions. If the community *over*commits to particular theories, science gets stuck.

But crucially, the existence of narrow points of view is not a problem. It's necessary. Any point of view is partial, but it doesn't mean we shouldn't have viewpoints. Even obsession can be powerful when we obsess on a problem at work and occasionally achieve good results.

Within the progress of science, temporary commitment to a paradigm is important for healthy functioning – Kuhn even argues that scientists should resist change, to a degree. In general, the point is not to shut down narrow concepts. The point is to limit them from becoming the sole and absolute determinants of behavior. In healthy functioning, different ideas are kept distinct but can also be called upon appropriately and related to one another (Gigerenzer and Gaissmaier, 2011; Hatano and Inagaki, 1984; Herzog and Hertwig, 2014; Tetlock, 1986).

2 Misalignment as overcommitment

'Growth for the sake of growth is the ideology of the cancer cell.'
Edward Abbey

We will return in Section 3 to more examples of overcommitment as the opposite of health, and there we will look at how living systems avoid overcommitment and maintain health. But with the beginnings of an operational definition of overcommitment, we now tie it to the AI alignment problem.

While people disagree on exactly what behaviors or properties of an AI system are aligned, there is general agreement that alignment means working toward futures that are healthy and flourishing. In the previous section we made a case that health and flourishing, in a very broad sense, are the opposite of overcommitment. Following this line of reasoning, we propose that alignment can be viewed as the problem of avoiding overcommitment to any particular form[2]. No matter how good or complete our current concepts or specifications or lenses are, setting them in stone as absolutes is not aligned. A system is misaligned if it has some fundamental cap on its sensitivity to a larger context.

To give a flavor of this mapping, we look at a few examples of alignment failures as problems of overcommitment. We'll then ask about the relationship between this broad concept of alignment and moral and normative notions of 'good' and 'should'. And finally we'll reflect on how large our definition makes the problem of alignment.

2.1 Optimizing for particular goals

It's difficult to specify what our values are, or the way we want the world to be. Any way we have of writing down or formalizing what we want misses important things. If we ask a powerful AI system to optimize for that formalization – in other words, to give us what we've said we want – the results are paradoxically disastrous (Amodei et al., 2016; Gabriel, 2020; Grossman and Hart, 1986; Hadfield-Menell and Hadfield, 2019; Krakovna et al., 2020; Russell, 2019; Wiener, 1960; Zhuang and Hadfield-Menell, 2020). Suppose an AI's objective is to increase humans' subjective experience of wellbeing. Under reasonable definitions, achieving this objective is most efficiently achieved by imprisoning humans and directly stimulating neurons to trigger

[2]Does this mean human welfare is indistinguishable from smallpox welfare? No. All the existing local perspectives in the world are vitally important. Some relativism is useful: for example, when it helps us appreciate the plurality of human values. But overcommitment to relativism is misaligned. AI comes into existence amid a profound network of existing reality which is saturated with meaning and importance. The point is to collaborate with all this form and structure, not to extinguish it.

our experience of wellbeing (Bostrom, 2014). Doing too good a job of optimizing for any formalized goal is misaligned by being overcommitted to the myopic form of that goal.

2.2 Concentration of human power

As a second clear example of misalignment-as-overcommitment, AI potentially conveys immense power to those who control it. In some scenarios, a small number of humans will have the majority of control over AI systems, facilitating dominance over other humans. These scenarios appear more likely as the persuasive power of technology increases (Costello et al., 2024; Hackenburg et al., 2025; Woolley and Howard, 2018), autonomous weapons place lethal force in a small number of hands (Scharre, 2018), surveillance and analytics improve, and the need for human labor decreases (Drago and Laine, 2025; Ford, 2015; Susskind, 2020). Concentration of human power overcommits to the goals and interests of a few individuals, at the expense of others.

2.3 Conceptual monoculture

Conceptual monoculture is overcommitment to particular beliefs, ideas, frames, values, problem-solving approaches – loss of diversity across a group. In many kinds of systems, monoculture creates fragility and leads to lower performance of the system as a whole (Haldane, 2013; Kleinberg and Raghavan, 2021; Scott, 1998; Tilman, 1996).

Current AI systems draw from a conceptual manifold that is, at least in some ways, impoverished relative to humans (Crawford, 2021; Kirk et al., 2023; Messeri and Crockett, 2024; Selwyn, 2024). Recent studies have discovered that while individual AI outputs are typically judged as superior to human outputs, the AI outputs are also more homogenous (Agarwal et al., 2025; Beguš, 2024; Doshi and Hauser, 2024; Kosmyna et al., 2025; Padmakumar and He, 2023; Xu et al., 2025; Zhou and Lee, 2024).

This narrow manifold might get broadcast to the whole world. At least a billion people around the world now use AI for everything from relationship advice to industrial maintenance (CCIA Research Center, 2025; Chatterji et al., 2025; Honeywell, 2024; McCain et al., 2025; OpenAI, 2025; Singla et al., 2025; TechCrunch, 2025). Yet because frontier models are difficult and expensive to produce, the massive usage is routed through a handful of models (Bommasani, 2021).

If centralized AI models broadcast their lower-diversity concepts to the whole world, there's a risk of global decrease in diversity. Since humans are both influenced by AI and a source of training data, the homogenization could even become recursive (Chaney et al., 2018). However, it is worth noting that none of the studies cited here made a best effort attempt to use AI systems in a more thoughtful way that could increase rather than decrease diversity. This gives hope, which we return to in the final section.

2.4 Instruction following

Failing to follow instructions probably most often also reflects lack of sensitivity to context, i.e., overcommitment to myopic patterns. Suppose the user asks an AI chatbot to write a poem about an elephant, and the AI instead writes a poem about a giraffe. With the chatbots of today, we can be fairly confident the cause of this divergence is not an internal spark of life where the AI system wisely decided a giraffe poem would be more apropos. Instead, the cause is probably a collapse of context. For example, models are prone to 'shortcut learning' where they over-rely on superficial correlations (even a single keyword) (Geirhos et al., 2020); they get fixated on data that was overrepresented in training (Reynolds and McDonell, 2021; Xu et al., 2024; Zhao et al., 2021); they lack flexibility in attending to the right positions in their input (Liu et al., 2024). Today, we are still in the regime where making AI systems more responsive to human instructions usually involves more subtlety, more sensitivity and less overcommitment.

There are important exceptions, however: we generally don't want AI systems to follow harmful instructions. When the AI system correctly refuses harmful requests, it is applying its own context to avoid overcommitment to human instructions. In these situations, the AI's designers have effectively decided there's a risk that the user is not fully sensitive to longer-sighted implications of their request. By extrapolation, as AI systems continue to gain scope, we should expect less direct compliance with human instructions (Bostrom, 2014; Hadfield-Menell et al., 2016; Milli et al., 2017; Russell, 2019; Yudkowsky, 2004). Rather than literally fulfilling a request, there might be a better response which achieves a deeper unstated intent of the user, or achieves an outcome aligned with the interests of more people or the longer-term future.

2.5 Normativity and human values

We've described a concept of alignment that deviates a bit from standard definitions. How is this concept related to human values, moral concepts of good, or normative ideas of what an AI ought to do?

The most straightforward notions of values or morals anchor on what we can relatively easily express. This kind of value might include improving subjective wellbeing for humans, reducing suffering or minimizing inequality, in ways that can be operationalized and measured. They are formalizable or close to formalizable.

However, values cast in that way are not very satisfying. As we described above, when values are formalized, they are vulnerable to proxy failure (John et al., 2023; Kurth-Nelson et al., 2023). If we think we've written down what we think we value, and then someone does a good enough job giving us the thing we said we want, the outcome is inevitably harmful in a broader sense.

One way to robustify values is allowing them to include things that are difficult to express formally (Dreyfus, 1972; Nussbaum, 2001; Polanyi, 1966; Scott, 1998; Varela et al., 1991; Wittgenstein, 1922). This kind of value might stretch far below language into subtle, contextual intuition that involves our bodies, communities and natural environment. Another extension is to allow values that are continually evolving in an open-ended way (Dewey, 1939; Gadamer, 1960; Murdoch and Midgley, 2013; Nietzsche, 1883; Singer, 1981; Williams, 1985). These values change as we ourselves continue to develop and evolve. Any concepts we have about them at any given point in time are inevitably incomplete, just like a planarian doesn't have the concepts to entertain the kinds of values we talk about today. The resistance of values to being fully captured by language or concepts might be something we value – in a way that is itself changing. Even the concept of 'values' is a form we might over-index on.

2.6 Overcommitment to any form is misaligned

'Truth, like love and sleep, resents
approaches that are too intense.'
W. H. Auden

Let's recap how big the problem is. AI safety researchers have identified many particular versions of overcommitment and developed or proposed solutions for them. For example, concentration of power might be mitigated by democratic oversight and involvement of more people in AI design decisions (Birhane et al., 2022; Dafoe, 2018; Lazar and Nelson, 2023; OpenAI, 2023; Selbst et al., 2019; Sloane et al., 2022); or through redistribution mechanisms (Gough, 2019; O'Keefe et al., 2020; Sharp et al., 2025; Susskind, 2020). Value lock-in might be mitigated by improving our mechanistic understanding of AI systems so we can, for example, detect and correct the systems if they develop hidden ways of resisting our efforts to change their goals (Anthropic Research Team, 2024; Bereska and Gavves, 2024; Burns et al., 2022; Olah et al., 2020); or by designing AI systems that want to obey human preferences but treat these preferences as something uncertain that must be learned (Hadfield-Menell et al., 2016, 2017; Jeon et al., 2020; Russell, 2019; Shah et al., 2020).

But there is a deeper problem. Any conceptual scheme, taken too seriously, is misaligned; therefore, no particular approach can achieve alignment[3]. An AI system could overcommit to the language for describing the space goals and values live in (Bobu et al., 2020; Soares and Fallenstein, 2014), to an algorithm for learning human preferences, to our concepts of agency or representation, or even to concepts we currently use but can't see because they are tautological to us. This problem can be viewed as a generalization of proxy failure (John et al., 2023) or generalization of the outer alignment problem (Hubinger et al., 2019). It's not only particular objectives that are subject to overcommitment failures, but any form at all, including what we ourselves unconsciously hold as axiomatic.

In the past, humanity has always iterated on technological solutions which, at any given moment, have imperfect forms. But AI poses a special kind of risk, because that iterating process might not work as it has in the past. AI already has some remarkable properties, such as rapid global adoption, intensive use of resources, concentration of information flow and human use patterns (such as anthropomorphization and offloading a large swath of cognitive activity). More speculatively, in the future AI may exhibit superhuman intelligence and recursive self-improvement, without being limited to a restrictive biological substrate. These distinctive properties may create more difficulty in iterating on imperfect solutions, compared to past technologies (Bostrom, 2014). Indeed, it's been suggested that AI explains the Fermi paradox – the puzzle of why we don't see signs of intelligent life anywhere else in the universe (Bostrom, 2008; Garrett, 2024). To use the language of this paper, as civilizations become intelligent, they develop the capacity to give themselves the myopic form of what they think they want. If that capacity develops faster than boundaries that contextualize it, it may lead many civilizations to overcommitment and collapse.

[3]Of course, this does not mean we should have no scheme. Quite the opposite. Throwing away schemes capriciously is just as over-fixated as any other particular form.

3 Contextualization and living boundaries

'When forced to work within a strict framework, the imagination is taxed to its utmost–and will produce its richest ideas. Given total freedom the work is likely to sprawl.'
TS Eliot

'Nature's imagination is so much greater than man's, she's never going to let us relax.'
Richard Feynman

To start to think about an answer to the big problem, let's look again at life for inspiration. How does it maintain its intricate multiscale balance and open-ended trajectory (Stanley et al., 2017) despite the constant threat of collapse into one overcommitment or another?

In this section, we examine how living systems, including human systems, function as an extraordinarily nuanced web of semi-permeable boundaries. Semi-permeable boundaries contextualize forms as part of larger systems by allowing them to interact with other forms without one part dominating or the system collapsing to homogeneity. Life is both the contextualizer and the thing being contextualized. For example, cognitive control allows a particular impulse to exist usefully as a flexible part of the organism's whole motivation. Contextualization means that an entity's distinctiveness is preserved (or even enhanced), yet it no longer exists as an absolute or in isolation but rather within a larger context. Individual forms thrive, but the larger system accommodates partial perspectives rather than locking in a single story of what matters.

The boundaries of life carry a remarkable amount of information. The life that exists today is exactly what has successfully traced a path through time along a remarkable knife edge, avoiding collapse into either excess or insufficient stability (Bak, 2013; Kauffman, 1992; Kirchhoff et al., 2018; Prigogine and Stengers, 1984; Schrödinger, 1944). Along this path, the forms of life have become fractally complex with each part carrying traces of the many contexts it has participated in. This kaleidoscope, now including humans' storage and exchange of mental patterns, are interrelated through shared heritage, ecological interactions and facing the same world.

3.1 Law

'Unity without uniformity and diversity without fragmentation.'
Kofi Annan

Individual actors in a society and in an economy each act from their own perspective. Each actor's perspective is myopic. Myopia does not always mean selfishness in the sense of valuing only one's own wealth or physical wellbeing (Becker, 1974; Crockett et al., 2014; Henrich et al., 2001). But an actor cannot know everything or fully understand the motives and beliefs of others.

Without boundaries, social systems tend to overcommit to one actor's perspective or interests. This domination results in collapse and an impoverished system. For example, a company's profit motive, if unresisted, leads to suppression of competition, deception, and exploitation of individuals (Bakan, 2006; Baran, 1966; Dalrymple, 2019; Goldacre, 2014; Smith, 1776). An individual's desire for power and social dominance can lead to disempowering or silencing of others and even direct infringement on the autonomy and wellbeing of others (Hawley, 2003;

Sidanius and Pratto, 2001; Tepper, 2000). Even genuinely held, ostensibly prosocial beliefs lead to conflict and suppression when different groups have different perspectives (Greene, 2013; Haidt, 2012; Scott, 1998).

Law, when it works well, is a boundary against dominance of any actor's motives. A person is motivated by a dispute to kill another person, but the law forbids murder. A business tries to maximize its success, but the law bans environmental exploitation, false advertising, and anti-competitive practice.

Effective laws do not annul the myopic drives of particular actors, but rather *contextualize* them within a larger system. Under ideal circumstances, the boundary of the law reroutes the energy of a myopic drive in more productive direction. A would-be murderer, unwilling to face the penalty of the law, might seek a dispute resolution establishing a stable framework that supports future prospering of both parties. A business wanting to expand, but constrained to act within the law, is driven to build better products (Ambec et al., 2013; Ashford et al., 1985; Wu, 2011).

Intelligent agents do not necessarily accept boundaries set on their desires. The law must adapt as its loopholes are discovered. Like other systems in the living world, it forms an evolving network of boundaries (Burns and Kedia, 2006; Campbell, 1979; Kerr, 1975; Ordóñez et al., 2009). The evolving laws gradually acquire grounded wisdom as they are tested against many different situations and motives.

3.2 Problem solving in groups

'I could also observe, time and again, how too deep an immersion in the math literature tended to stifle creativity.'
Jean Écalle

'There's more exchange of information than ever. What I don't like about the exchange of information is, I think that the removal of struggle to get that information creates bad cooking.'
David Chang

In 1968, the nuclear submarine USS Scorpion vanished en route from the Mediterranean to Virginia (Craven, 2002; Sontag et al., 1998; Surowiecki, 2005). The Navy started a search, but the amount of ocean where the vessel could be was enormous. John Craven, Chief Scientist of the U.S. Navy's Special Projects Office, devised an unusual search strategy. He assembled a diverse group of mathematicians, submarine specialists, and salvage operators. But he didn't let them communicate with each other. Each expert had to use their own methods to come up with their own estimate of where the Scorpion should be. Craven then aggregated the independent estimates into a single prediction. Astonishingly, the wreckage was found only 220 yards from this spot.

When solving problems, different people bring different perspectives and approaches. Each method processes the available data using a different toolkit. Under favorable conditions, combining the approaches of multiple contributors yields better results than any individual working alone. This 'wisdom of crowds' effect has been documented in numerous domains of problem solving (Condorcet, 1785; Surowiecki, 2005).

However, there is a perpetual danger of overcommitment. The wisdom of crowds is diminished if a group lacks diversity, either ab initio or as a result of within-group communication and influence (Hogarth, 1978; Hong and Page, 2004; Ladha, 1992; Surowiecki, 2005). Controlled experiments, as well as analyses of key decision moments in real groups, find that groups collectively reach irrational or suboptimal solutions when diverse and dissenting viewpoints are lost to a narrower set of ideas (Anderson and Holt, 1997; Becker et al., 2017; Bernstein et al., 2018; Diehl and Stroebe, 1987; Flowers, 1977; Frey and Van de Rijt, 2021; Janis, 1972; Stasser and Titus, 1985). Unstructured communication methods like open discussion have a special vulnerability of rhetorical force dominating over epistemic merit. At the same time, sharing information is essential for the benefits of group wisdom and cooperative behavior. There is therefore a tension between overcommunication where diversity is lost and undercommunication where diversity is not leveraged.

The crux is semi-permeable boundaries: wisely transmitting the right information at the right time, in the right way. Thoughtful strategies for communication are like transmembrane channels that allow the right molecules in and out of the cell at the right time. They protect and enhance diverse problem solving approaches while also allowing productive interaction between them. Semi-permeable boundaries are contextualizing: they retain individuality while also situating it within relationships to other entities.

Many varieties of semi-permeable boundary are effective in boosting group performance, including: creating decentralized topologies where group members only communicate with nearby neighbors (Becker et al., 2017; Mason et al., 2008); defining rules that incentivize acting according to one's own belief rather than following the crowd (Bazazi et al., 2019; Hung and Plott, 2001); modeling the strengths and weaknesses of each group member (Welinder et al., 2010); promoting leadership styles where one person's views are less likely to dominate (Flowers, 1977; Leana, 1985); and periodically breaking up into subgroups or rotating membership (Baron, 2005; Bebchuk and Cohen, 2005; Feldman, 1994; Hauer et al., 2021; Janis, 1972; Kane et al., 2005; Owen, 2019; Straus et al., 2011; Sutton and Louis, 1987; Trainer et al., 2020; Vafeas, 2003; Wu et al., 2022). In a later section, we will look at boundaries within an individual, such as skepticism, that make it easier to interact with others without overwriting one's own beliefs.

The importance of balancing communication with independence also shows up in controlled experiments. Frey and Van de Rijt (2021) elicited votes about general knowledge questions (such as, 'In which year did Germany invade Denmark?') from a group. In one condition of the experiment, participants voted sequentially and could see the running tallies of previous voters. Compared to independent voting, final group means were less accurate in the sequential condition, because early mistakes got baked in to the group's belief. In a related experiment, Bernstein et al. (2018) tasked small groups with solving instances of the traveling salesman problem. Groups that exchanged information continuously tended to reach poor final outcomes, compared to groups with less communication. When one individual discovered a solution that looked compelling but was actually a dead-end, other group members collapsed on this dead-end and the group as a whole made less progress.

3.3 Cells

'It is by avoiding the rapid decay into the inert state of equilibrium that an organism appears so enigmatic.'
Erwin Schrödinger

One of the most reified examples of a boundary in nature is the cell membrane (Alberts et al., 2022; Bray, 2019; Harold, 2001; Lane, 2015; Watson, 2015). Without the membrane, the pressure of chemical gradients would rapidly homogenize the cell's contents with the outside – severe overcommitment to a uniform state, collapsing the subtlety of the cell's structure. Thanks to the membrane, both the cell and the outside can exist, a more diverse, less symmetric arrangement (Anderson, 1972; Prigogine and Stengers, 1984; Schrödinger, 1944; Turing, 1952).

Cell membranes are semi-permeable: they prevent the conditions outside from grossly overwriting the inside, but they do not block interactions wholesale. Via the sophistication of the membrane, outside information is selectively gated and transformed. To maintain its semi-fragile internal state between stasis and randomness, the cell needs a constant influx of energy. Channels permit certain small molecules to enter but not others, and these permissions are switched on and off according to momentary context. Endocytosis brings larger structures from outside into the cell. Cell surface receptors, when activated by external ligands, initiate intracellular signaling cascades that little resemble the ligand: an even more heavily curated form of influence. These and other processes allow information from the outside to influence the inside – not in a totalitarian way but in a nuanced way, mediated by the intelligence of the boundary.

Semi-permeable boundaries store and put to work the potential energy of the asymmetry between different forms. The same gradients that could annihilate the cell to equilibrium instead drive useful signaling, like action potentials in nerve and muscle cells. Instead of short-circuiting, myopic forces are contextualized to propel the continuation of life.

In multicellular organisms, most of the 'outside' is defined by other cells. For organisms to work well as a whole, even though the cells are largely 'on the same team', it's important that they don't blend into each other. Neurons rely on this principle dramatically, stretching out long processes to almost touch other neurons but then leaving the gap of the synapse. Synapses allow the network to precisely isolate many separate signals and direct information to relevant cells. They boost computational power as the signals are gated and transformed, and the nature of this transformation is plastic, storing a huge amount of information. Symbiogenesis is another example of how cells achieve more by retaining some discreteness than by smoothly blending together (Margulis and Sagan, 1986, 1995).

Again, collapse or overcommitment is always relative. For example, programmed cell death is catastrophic collapse at the level of the dying cell, but it can be beneficial or even necessary for the organism the cell belongs to.

3.4 Genetic recombination

Sex is costly. An organism must find a mate in the vast and dangerous world, and half of the creatures can't reproduce (Goodenough and Heitman, 2014; Lehtonen et al., 2012; Maynard Smith, 1971, 1978). Yet nearly all eukaryotes reproduce sexually[4] (Bell, 1982; Speijer et al., 2015). This raises the question: what is so great about sex?

In asexually reproducing species, all descendants of an organism are nearly clones, up to mutations within the lineage. Being permanently locked together gives the genes strong influence on each other. Selection can't act on one gene without dragging on the others. For example, suppose there are two genotypes within an asexual population, carrying different alleles at each of two different loci, as a result of mutations. One of the loci is currently fitness-neutral while the other is subject to selection pressure. The selection pressure tends to cause one of these genotypes to outcompete the other, eliminating one variant at the neutral locus. In other words, tight linkage between genes puts direct downward pressure on genetic diversity (Charlesworth et al., 1993; Hudson and Kaplan, 1995). Additionally, if two different beneficial mutations arise in two different organisms, they compete with each other. The only way for a single organism to obtain both beneficial mutations is if one arises again within the subpopulation that already carries the other, which is unlikely and therefore slow (Crow and Kimura, 1965; Felsenstein, 1974; Fisher, 1930; Hill and Robertson, 1966; Muller, 1932; Weismann, 1889). Conversely, if a deleterious mutation arises, all of the other genes in that lineage are stuck with it forever – unless there is a reverse mutation, which is rare (Keightley and Otto, 2006; Muller, 1932). An asexual species has rigid rather than flexible interaction between genes: it overcommits to particular genetic arrangements.

Recombination is a boundary that softens the rigid interactions between genes. It frequently breaks up the relationships between genes, assembling them into new genomes, effectively saying, 'don't get overconfident in that genetic arrangement; hold each arrangement more lightly'. From a gene's point of view, this looks like, 'don't get overly dependent on specific other genes'. Aspects of the genome that work well are propagated, like sodium ions gated into a neuron during an action potential, and poorly-working aspects are discarded. Sex contextualizes genetic arrangements.

Boundaries encourage lightly-held, modular interactions. By not overcommitting to a particular genome, sex encourages genes to flexibly interact with other genes (Clune et al., 2013; Dawkins, 1976; Holland, 1975; Livnat et al., 2008, 2010; Wagner and Altenberg, 1996). Instead of being overfit to a particular context, genes develop a robust identity that's both independent and interfunctional. Recombination puts genes under pressure to evolve a generalized, grounded wisdom that reflects the structure of the world, like a person learning multiple languages and extracting the underlying commonalities. At the same time, because each gene is always operating in the presence of other genes, it develops its own distinct point of view that adds unique value to a genome.

3.5 Cognitive control

Cognitive control is a broad class of boundaries on particular drives, goals and strategies (Botvinick et al., 2001; Braver, 2012; Miller and Cohen, 2001; Miyake et al., 2000). In section 1.1, we looked at how organisms can overcommit – unhealthily – to particular motives or strategies for fulfilling the motives. Cognitive control contextualizes motives, strategies and thought patterns by allowing them to exist and perform useful functions without dominating.

[4]And all known species exhibit some kind of gene transfer, performing a related function.

For example, I may have a drive to consume food, but I can apply control to avoid overeating. I might work obsessively on a project while also having a rule that I must go to bed at 10 pm. This boundary doesn't block me from temporarily taking a strong perspective, but it does place contextual limits on it. Cognitive control, when functioning well, is a semi-permeable boundary: it situates myopic patterns within a larger system.

Control translates the pressure of motivation into higher-order structure. When nothing stops a particular drive or goal or strategy from dominating behavior, it tends to follow a shortest path defined under its own myopic understanding of the world. For example, the chickens in Section 1.1 wanted food and tried to take the shortest path toward it in the naive sense of a straight line through space. In the backwards world created by the experimenter, this action does not accomplish the deeper goal of reaching food, for which moving spatially toward food is only a proxy. The chicken's motivation is short-circuited: it expends energy without making progress on the deeper goal. Humans can easily solve the task by inhibiting their prepotent impulse to approach food. The boundary of control breaks the symmetry of congruent action. In general, semi-permeable boundaries promote formation of new structure by placing contextualizing limits.

3.6 Information in the brain

'Memory is not an average of experience.'
David Marr

The brain miraculously keeps many pieces of information distinct from one another. If you picture a highly connected network of neurons with their signals continually impinging on one another, it's not obvious that this would be an easy thing to accomplish. In this section, we review a selected handful of mechanisms by which the brain maintains semi-permeable boundaries between different signals. Each paragraph below focuses on one of these mechanisms. There are many more that we do not cover. The brain is perhaps the most extraordinary example in nature of a system of semi-permeable boundaries supporting the proliferation of multitudinous forms that develop their own richly distinct identities yet are also meaningfully linked together.

Lateral inhibition is a central tenant of neural organization (Douglas and Martin, 2004; Hubel and Wiesel, 1962; Isaacson and Scanziani, 2011). Lateral inhibition means the activity of a neuron is reduced when its neighbors are active. This segregates information to create and sustain distinct neural representations. Lateral inhibition was first studied in the nerve cells of the eye, where it enhances contrast at the edges of stimuli (Hartline et al., 1956). When a photoreceptor in the retina is activated by light, it sends signals forward toward the brain; but it also activates inhibitory interneurons, which suppress adjacent photoreceptors and their downstream targets. This amplifies the perception of borders and contours. And the same principle operates throughout the brain. In visual cortex, for example, inhibition sharpens selectivity of neurons for abstract visual features like the orientation of a line (Sillito, 1975).

The brain uses inhibition organized into oscillatory dynamics to keep memory items separated (Jensen and Mazaheri, 2010; Klimesch et al., 2007; Lisman and Jensen, 2013; Roux and Uhlhaas, 2014). Distinct items fire at different phases of the 8-12 Hz alpha oscillation. The inhibitory phase of the alpha rhythm silences all but one item at any given moment. By segregating firing in phase space, multiple memories are held simultaneously without interference.

The circuit architecture of hippocampus separates experiences or concepts into distinct representations, avoiding interference between similar memories (Colgin et al., 2008; Leutgeb et al., 2007; Marr, 1971; McClelland et al., 1995; McNaughton and Morris, 1987; Muller and Kubie, 1987; Treves and Rolls, 1994). Inputs from entorhinal cortex are distributed via mossy fibers to a much larger population of dentate gyrus granule cells, creating sparse, orthogonal codes in dentate gyrus. This way, situations or ideas that are superficially similar but functionally different are kept cleanly separated in neuronal activity space – a unique neural fingerprint for each distinct concept or memory. This prevents, for example, yesterday's memory of where you parked your car from interfering with today's memory of where you parked your car in the same parking ramp.

Compared to other animals, the human brain especially attempts to discretize its experience into approximately symbolic representations (Behrens et al., 2018; Dehaene et al., 2022; Smolensky, 1990; Touretzky and Hinton, 1988). The capacity to separate things into nearly-discrete entities and then recombine them in vast numbers of structured ways powers the extraordinary human capacity for reasoning (Chomsky, 1957; Fodor, 1975; Kurth-Nelson et al., 2023; Lake et al., 2015; Pinker, 1994). Semi-permable boundaries keep forms distinct while enabling them to flexibly and modularly interact. Like genes participating in many genomes, discretized neural representations participate in many structured combinations. This encourages each entity to develop an identity that both is distinct and also reflects a more generalized picture of the world.

More broadly, healthy brain dynamics live at a sweet spot between excessively stable synchronized patterns and chaotic uncorrelated noise (Bak et al., 1987; Beggs and Plenz, 2003; Chialvo, 2010; Deco et al., 2011; Haldeman and Beggs, 2005; Kotler et al., 2025; Rabinovich et al., 2008; Shew et al., 2011; Tognoli and Kelso, 2014). In this regime, the brain has access to a huge repertoire of patterns it can explore temporarily without overcommitting or getting stuck.

Loss of dynamic flexibility, where the brain's activity becomes more stereotyped and no longer explores as wide a repertoire of states, is tied to lower cognitive performance (Cocchi et al., 2017; Garrett et al., 2013; Grady and Garrett, 2014; Müller et al., 2025; Shew et al., 2009). More extreme stereotypy corresponds to severe dysfunction. For example, in Parkinson's disease, basal ganglia and cortical circuits collapse into excess synchrony and lose the flexibility needed to guide nuanced motor outputs (Brown, 2003; Hammond et al., 2007).

3.7 Interpersonal dynamics

'Stand together yet not too near together, as the oak tree and the cypress grow not in each other's shadow.'
Kahlil Gibran

Psychoanalysis introduced the concept of 'boundaries' in human psychology, distinguishing what is the self from what is outside or other (Federn, 1928; Tausk, 1919). Early works applied the concept to psychosis, where those boundaries were thought to be blurred. But the need for clear self-other boundaries was also thrown into relief by the intimacy of the therapeutic relationship. In complex internal territory, it became harder to disentangle which experiences really belonged to someone and which were attributed in imagination by the other person (Freud, 1894, 1910). Analysts risked harming patients by imposing their own beliefs and desires, even to the extent of sexual abuse or psychological domination (Gabbard and Lester, 1995).

The concept was enriched by Gestalt therapists, who agreed that boundaries can be too permeable; but added that they can also be too rigid, causing isolation and stagnation (Perls et al., 1951; Polster and Polster, 1974; Yontef, 1993). Family systems theorists and subsequent work further emphasized that lack of boundary in close relationships leads to enmeshment and loss of autonomy, while excessively rigid boundaries lead to isolation (Bowen, 1978; Brown, 2012; Cloud and Townsend, 1992; Minuchin, 1974). In attachment theory, people with an anxious attachment style struggle to set boundaries for fear of alienating others, while people with an avoidant attachment style develop overly rigid and isolating boundaries (Ainsworth et al., 1978). Strengthening the agency of the self through semi-permeable boundaries is foundational for psychological health: meaningful connection with other people while preserving integrity of the self.

As with other living systems, humans have a rich array of psychological boundaries, with intelligence in their nuance. Anger, historically often viewed as sinful and irrational, is now seen as part of our system of boundaries: an important signal that our integrity is being violated (Lerner, 1985; Sell, 2011; Videbeck, 2010). Healthy shame is suggested to operate as a bound on our own selfishness (Bradshaw, 1988). Some psychologists argue that the incest taboo reroutes desires, which would otherwise be short-circuited, into productive activity (Freud, 1913; Lévi-Strauss, 1949; Stein, 1973). Assertiveness forms a boundary against the drives of other individuals (Smith, 1985). Skepticism protects us from credulity and having our own experience overwritten by the assertions of others (Lewandowsky et al., 2012; Sperber et al., 2010). Boundaries take many forms and continue to evolve as we learn across our lifetime.

Without boundaries, interactions tend to result in one person being dominated by another: a patient's own beliefs replaced with those of an analyst, or the desires of one person in a relationship ignored. With semi-permeable boundaries, we have rich internal worlds. We are sensitive to each other, but there is also enough space for our internal experience to flourish without being immediately overwritten by external signals. Our internal experience is contextualized in relationship to other individuals, creating new structure: mutual understandings, relationships, communities, cultures. And as individuals we grow as we are shaped by different contexts. This metastability or loose coupling is interestingly reminiscent of brain dynamics.

3.8 Contemplative practice

'The world is perfect as it is, including my desire to change it.'
Ram Dass

Awareness is contextualization. Think of an assumption somebody has that's never been questioned. That assumption could be lifelong and self-defining, or it could be fleeting and perceptual, like the assumption that the thing I'm touching is a keyboard. Within its own frame, it has a kind of tautological truth, a near-absolute formality. But sometimes there's a moment of stepping back, where the assumed form becomes an object in awareness, and the assumption is contextualized. The person realizes it's not an absolute truth standing alone, but rather a form in their mind.

We inevitably get stuck because we have assumptions about things we tautologically can't allow to happen – things that are Bad. What we do to prevent the bad thing from happening is what causes the bad thing to continue. Our collapsed patterns hold the tension that paradoxically creates the unease they resist. Awareness contextualizes these dynamics. It feel infinitely scary from the original frame, because it's allowing the bad thing. But from the new frame, the bad thing is just another content of experience. The fear or wrongness of is no longer an absolute but exists in context. So awareness brings healing and growth.

Awareness is an evolving system of boundaries: it limits overcommitment to any thing. The boundaries are semi-permeable: becoming aware of a belief doesn't make the belief wrong in an absolute sense any more than it was right in an absolute sense. Awareness holds us at the knife's edge of not collapsing exclusively into any particular forms. This activates a deeper sensitivity to our own livingness and to the world. Subtler forms, which would have been erased by overcommitment to other forms, instead play a role in a richer overall internal structure. Our own potential within the world creatively emerges in continued newness.

The whole idea of awareness is, of course, incomplete too. Once we picture awareness as an object, it's not the thing we're talking about. By construction, contextualization is an unsolvable mystery from any particular point of view.

The orientation toward not overcommiting to particular forms within experience is familiar in art, poetry, music, dance. The meaning of art is open-ended and changes with context – it has an inner life. What we value is perhaps something about the subtlety and the resistance art has to being pinned down into a formalism. It moves us.

4 An aligned future

'We can love the beautiful, and believe in it, and thereby open ourselves to an understanding of love that does not dominate, but cherishes the independence and beauty of the loved.'
Martha Nussbaum

What does the opposite of overcommitment look like in a future co-created by AI? In living systems, evolving semi-permeable boundaries contextualize partial forms to be more long-sighted in time and space, increasing subtlety and potential. Now we take a preliminary look at applying what we learn from life to create an aligned future.

4.1 Boundaries

The use of boundaries against overcommitment in technology is as old as technology. As soon as we started building things, we had to build in boundaries, because we want the things we build to be robust and useful and not to collapse into degenerate states. We put circuit breakers in the power grid, governors in steam engines, escapements in clocks, ReLUs in neural networks. In modern AI research we have personalization through the context, access to user data, on-device learning. We have conditional computation creating separation between parts within a large model. We have dropout, cross-validation, causal masking. Many safety methods are boundaries, including safety post-training, guardrail models, red teaming, mechanistic interpretability, government oversight and so on (Gabriel et al., 2025). Increasingly critical are ongoing evaluation and monitoring of deployed AI systems (Grey and Segerie, 2025; Myllyaho et al., 2021; Yampolskiy, 2025). There is increasing awareness that over-attachment to fixed optimization targets is often counterproductive (Kumar et al., 2025; Stanley and Lehman, 2015; Stanley et al., 2017).

At a social level, we have labs taking different approaches, nations with different cultures and strategic interests, ideas drawn from diverse fields like neuroscience and physics. Interestingly, the nature of this diversity has evolved rapidly in recent decades, as information exchange has improved with telecommunications, the internet, and now AI itself. One open question is how to preserve and enhance diversity in human culture and concepts, at scales anywhere from national or ethnic groups to individuals, and how to productively put the diverse elements in contact with each other. Because we can now use AI to effectively get answers from the rest of humanity instantly, an important aspect of this question is how we will structure our own use of AI to maintain our autonomy and diversity.

4.2 An ongoing process

But we hope our perspective also hints at something more. Even the most foundational assumptions are probably not final answers. The lenses we use to look at the world keep changing. True alignment is a process where whatever was previously axiomatic becomes a contextualized object. The point of alignment is not to say that any particular perspective is absolutely wrong or right. An aligned future will include continual reinvention of whatever concepts we have, including to the assumptions those concepts are built on, and the assumptions those assumptions are built on. Whatever concepts we currently have do not place hard limits on the future. Even the concept of 'not being overly attached to our concepts' is itself something we can release into contextualization.

As humans – whether AI researchers or any participants in social systems – this can be a mundane practical process of holding ideas with some skepticism, having patience to look at different timescales, listening to an internal voice of wisdom, entertaining conflicting perspectives. It can also be a profound process of self-awareness and personal growth. We continually evolve what we believe, even our self-definition, releasing beliefs into larger awareness without losing or erasing them.

Perhaps the most interesting question is this: what does the process of open-ended contextualization look like within AI systems, or in human-AI relationships? Is there a version of AI that continually contextualizes its own processes as partial truths? Do existing AI systems already do this to some degree, as they train and as they learn from interactions with humans and the world? What would it mean to take this process farther, for AI to continually release from exclusive attachment to any particular form? What can we do now to protect the potential for even any form of that releasing process to not be a final answer?

4.3 Conclusion

What we learn from living systems is that health or flourishing is not any particular form or concept. Therefore, alignment is not picking the right values or principles, or even the right system for learning them. It is not any method for interpretability or corrigibility. All of these can be useful parts of alignment. But alignment itself is the continued dance of contextualizing any particular form, creating deeper relationship with the neverending mystery of the world. In this way, AI can participate in intense flourishing of an evolving world even far beyond current human conceptualization.

5 Acknowledgements

Clark Potter for planting these ideas more than a decade ago. Zach Duer for comments on the manuscript.

6 Competing Interests

The authors declare no competing interests.

References

R. A. Adams, K. E. Stephan, H. R. Brown, C. D. Frith, and K. J. Friston. The computational anatomy of psychosis. *Frontiers in psychiatry*, 4:47, 2013.

T. W. Adorno, E. Frenkel-Brunswik, D. J. Levinson, and R. N. Sanford. *The Authoritarian Personality*. Harper & Brothers, New York, 1950.

D. Agarwal, M. Naaman, and A. Vashistha. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2025.

M. D. S. Ainsworth, M. C. Blehar, E. Waters, and S. Wall. *Patterns of attachment: A psychological study of the strange situation*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1978.

B. Alberts, R. Heald, A. Johnson, D. Morgan, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell: seventh international student edition with registration card*. WW Norton & Company, 2022.

S. Ambec, M. A. Cohen, S. Elgie, and P. Lanoie. The porter hypothesis at 20: can environmental regulation enhance innovation and competitiveness? *Review of environmental economics and policy*, 7(1):2–28, 2013.

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

L. R. Anderson and C. A. Holt. Information cascades in the laboratory. *The American economic review*, pages 847–862, 1997.

P. W. Anderson. More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047):393–396, Aug 1972. doi: 10.1126/science.177.4047.393.

Anthropic Research Team. Mapping the mind of a large language model. Online research report / blog post, 2024.

APA. *Diagnostic and statistical manual of mental disorders*. American psychiatric association, 2013.

N. A. Ashford, C. Ayers, and R. F. Stone. Using regulation to change the market for innovation. *Harv. Envtl. L. Rev.*, 9:419, 1985.

A. Baddeley. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11):417–423, 2000.

P. Bak. *How nature works: the science of self-organized criticality*. Springer Science & Business Media, 2013.

P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: An explanation of 1/f noise. *Physical Review Letters*, 59:381–384, 1987. doi: 10.1103/PhysRevLett.59.381.

J. Bakan. The corporation. the pathological pursuit of profit and power, 2006.

T. Ballard, J. B. Vancouver, and A. Neal. On the pursuit of multiple goals with different deadlines. *Journal of Applied Psychology*, 103(11):1242, 2018.

B. W. Balleine, M. R. Delgado, and O. Hikosaka. The role of the dorsal striatum in reward and decision-making. *Journal of Neuroscience*, 27(31):8161–8165, 2007.

P. A. Baran. *Monopoly capital*. NYU Press, 1966.

R. S. Baron. So right it's wrong: Groupthink and the ubiquitous nature of polarized group decision making. *Advances in experimental social psychology*, 37(2):219–253, 2005.

G. Bateson. *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. Chandler Publishing Company, San Francisco, 1972. ISBN 0810204479.

S. Bazazi, J. von Zimmermann, B. Bahrami, and D. Richardson. Self-serving incentives impair collective decisions by increasing conformity. *PloS one*, 14(11):e0224725, 2019.

L. A. Bebchuk and A. Cohen. The costs of entrenched boards. *Journal of financial economics*, 78(2):409–433, 2005.

J. S. Beck, A. Beck, and J. Beck. Cognitive behavior therapy: basics and beyond. ed. *New York*, 2011.

G. S. Becker. A theory of social interactions. *Journal of political economy*, 82(6):1063–1093, 1974.

J. Becker, D. Brackbill, and D. Centola. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the national academy of sciences*, 114(26):E5070–E5076, 2017.

J. M. Beggs and D. Plenz. Neuronal avalanches in neocortical circuits. *Journal of neuroscience*, 23(35):11167–11177, 2003.

N. Beguš. Experimental narratives: A comparison of human crowdsourced storytelling and ai storytelling. *Humanities and Social Sciences Communications*, 11(1):1–22, 2024.

T. E. Behrens, T. H. Muller, J. C. Whittington, S. Mark, A. B. Baram, K. L. Stachenfeld, and Z. Kurth-Nelson. What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018.

G. Bell. *The Masterpiece of Nature: The Evolution and Genetics of Sexuality*. University of California Press, Berkeley, 1982. ISBN 978-0520045835.

L. Bereska and E. Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

E. Bernstein, J. Shore, and D. Lazer. How intermittent breaks in interaction improve collective intelligence. *Proceedings of the National Academy of Sciences*, 115(35):8734–8739, 2018.

A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, and S. Mohamed. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8, 2022.

A. Bobu, A. Bajcsy, J. F. Fisac, S. Deglurkar, and A. D. Dragan. Quantifying hypothesis space misspecification in learning from human–robot demonstrations and physical corrections. *IEEE Transactions on Robotics*, 36(3):835–854, 2020.

R. Bommasani. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

N. Bostrom. Where are they? why i hope the search for extraterrestrial life finds nothing. *MIT Technology Review*, pages 72–77, May 2008. URL `https://nickbostrom.com/papers/where-are-they/`. Originally published in the May/June issue.

N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. ISBN 0199678111.

M. M. Botvinick, T. S. Braver, D. M. Barch, C. S. Carter, and J. D. Cohen. Conflict monitoring and cognitive control. *Psychological review*, 108(3):624, 2001.

M. Bowen. *Family therapy in clinical practice*. Jason Aronson, 1978.

G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356): 791–799, 1976.

J. Bradshaw. *Healing the shame that binds you*. Health Communications, Inc., 1988.

T. S. Braver. The variable nature of cognitive control: a dual mechanisms framework. *Trends in cognitive sciences*, 16(2):106–113, 2012.

D. Bray. *Wetware: a computer in every living cell*. Yale University Press, 2019.

B. Brown. *Daring Greatly: How the Courage to Be Vulnerable Transforms the Way We Live, Love, Parent, and Lead*. Gotham Books, New York, NY, 2012. ISBN 9781592407330.

P. Brown. A rhythmic mechanism for communication in the cortex. *Trends in neurosciences*, 26(5):232–233, 2003.

C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

N. Burns and S. Kedia. The impact of performance-based compensation on misreporting. *Journal of financial economics*, 79(1):35–67, 2006.

D. T. Campbell. Assessing the impact of planned social change. *Evaluation and program planning*, 2(1):67–90, 1979.

G. Canguilhem. *Le normal et le pathologique*. Presses Universitaires de France, Paris, 1966.

R. N. Cardinal, J. A. Parkinson, J. Hall, and B. J. Everitt. Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience & Biobehavioral Reviews*, 26(3):321–352, 2002.

P. Carnes. *Out of the shadows: Understanding sexual addiction*. Hazelden Publishing, 2001.

CCIA Research Center. 2025 survey of product impact in the connected economy: Artificial intelligence. Spice ai report, Computer & Communications Industry Association, Nov. 2025. URL `https://ccianet.org/research/reports/2025-survey-of-product-impact-in-the-connected-economy-artificial-intelligence/`. Accessed: 2025-12-08.

G. Ceballos, P. R. Ehrlich, A. D. Barnosky, A. García, R. M. Pringle, and T. M. Palmer. Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science advances*, 1(5):e1400253, 2015.

A. J. Chaney, B. M. Stewart, and B. E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems*, pages 224–232, 2018.

B. Charlesworth, M. T. Morgan, and D. Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, 1993.

A. Chatterji, T. Cunningham, D. Deming, Z. Hitzig, C. Ong, C. Shan, and K. Wadman. How people use ChatGPT. Technical report, OpenAI, Sept. 2025. URL `https://cdn.openai.com/pdf/a253471f-8260-40c6-a2cc-aa93fe9f142e/economic-research-chatgpt-usage-paper.pdf`.

D. R. Chialvo. Emergent complex neural dynamics. *Nature physics*, 6(10):744–750, 2010.

N. Chomsky. *Syntactic Structures*. Mouton de Gruyter, The Hague, 1957.

H. Cloud and J. Townsend. *Boundaries: When to Say Yes, How to Say No to Take Control of Your Life*. Zondervan, Grand Rapids, MI, 1992.

J. Clune, J.-B. Mouret, and H. Lipson. The evolutionary origins of modularity. *Proceedings of the Royal Society b: Biological sciences*, 280(1755):20122863, 2013.

L. Cocchi, L. L. Gollo, A. Zalesky, and M. Breakspear. Criticality in the brain: A synthesis of neurobiology, models and cognition. *Progress in neurobiology*, 158:132–152, 2017.

L. L. Colgin, T. Denninger, M. Fyhn, T. Hafting, T. Bonnevie, O. Jensen, M.-B. Moser, and E. I. Moser. Frequency of gamma oscillations routes flow of information in the hippocampus. *Nature*, 455:125–129, 2008. doi: 10.1038/nature07278.

M. Condorcet. *Essai sur l'Application de l'Analyse a la Probabilité des Décisions Rendues a la Pluralité des Voix*. Imprimerie Royale, Paris, 1785.

T. H. Costello, G. Pennycook, and D. G. Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 2024.

J. P. Craven. *The Silent War: The Cold War Battle Beneath the Sea*. Simon and Schuster, 2002.

K. Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

M. J. Crockett, Z. Kurth-Nelson, J. Z. Siegel, P. Dayan, and R. J. Dolan. Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, 111(48):17320–17325, 2014.

J. F. Crow and M. Kimura. Evolution in sexual and asexual populations. *The American Naturalist*, 99(909):439–450, 1965. doi: 10.1086/282389.

J. P. Crutchfield. The calculi of emergence: computation, dynamics and induction. *Physica D: Nonlinear Phenomena*, 75(1-3):11–54, 1994.

A. Dafoe. Ai governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, 1442:1443, 2018.

W. Dalrymple. *The Anarchy: The Relentless Rise of the East India Company*. Bloomsbury Publishing, 2019. ISBN 9781408864401. URL `https://books.google.co.uk/books?id=-T2lDwAAQBAJ`.

R. Dawkins. *The Selfish Gene*. Oxford University Press, Oxford, 1976.

P. Dayan, Y. Niv, B. Seymour, and N. D. Daw. The misbehavior of value and the discipline of the will. *Neural networks*, 19(8):1153–1160, 2006.

E. De Bono. Lateral thinking. *New York*, page 70, 1970.

G. Deco and V. K. Jirsa. Ongoing cortical activity at rest: criticality, multistability, and ghost attractors. *Journal of Neuroscience*, 32(10):3366–3375, 2012.

G. Deco, V. K. Jirsa, and A. R. McIntosh. Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nature Reviews Neuroscience*, 12(1):43–56, 2011.

S. Dehaene. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking, New York, 2014.

S. Dehaene, F. Al Roumi, Y. Lakretz, S. Planton, and M. Sablé-Meyer. Symbols and mental programs: a hypothesis about human singularity. *Trends in Cognitive Sciences*, 26(9):751–766, 2022.

J. Dewey. Theory of valuation. *International encyclopedia of unified science*, 1939.

M. Diehl and W. Stroebe. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of personality and social psychology*, 53(3):497, 1987.

A. R. Doshi and O. P. Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290, 2024.

R. J. Douglas and K. A. Martin. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.*, 27 (1):419–451, 2004.

L. Drago and R. Laine. Defining the intelligence curse. `https://intelligence-curse.ai/defining/`, April 2025. Accessed: 2025-11-05.

H. L. Dreyfus. *What Computers Can't Do: The Limits of Artificial Intelligence*. Harper & Row, New York, NY, 1972.

K. Duncker. On problem-solving. *Psychological Monographs*, 58, 1945.

J. L. Evenden. Varieties of impulsivity. *Psychopharmacology*, 146(4):348–361, 1999.

P. Federn. Narcissism in the structure of the ego. *The International Journal of Psycho-Analysis*, 9:401, 1928.

D. C. Feldman. Who's socializing whom? the impact of socializing newcomers on insiders, work groups, and organizations. *Human Resource Management Review*, 4(3):213–233, 1994.

J. Felsenstein. The evolutionary advantage of recombination. *Genetics*, 78(2):737–756, 1974. doi: 10.1093/genetics/78.2.737.

Y.-J. Feng, D. C. Blackburn, D. Liang, D. M. Hillis, D. B. Wake, D. C. Cannatella, and P. Zhang. Phylogenomics reveals rapid, simultaneous diversification of three major clades of gondwanan frogs at the cretaceous–paleogene boundary. *Proceedings of the national Academy of Sciences*, 114(29):E5864–E5870, 2017.

P. K. Feyerabend. *Against Method*. Verso, 1975.

R. A. Fisher. *The Genetical Theory of Natural Selection*. The Clarendon Press, Oxford, 1930.

M. L. Flowers. A laboratory test of some implications of janis's groupthink hypothesis. *Journal of Personality and Social Psychology*, 35(12):888, 1977.

J. A. Fodor. *The Language of Thought*. Harvard University Press, Cambridge, MA, 1975.

M. Ford. *Rise of the Robots: Technology and the Threat of a Jobless Future*. Basic Books, New York, 2015.

M. J. Frank and E. D. Claus. Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological review*, 113(2):300, 2006.

A. Freud. *Das Ich und die Abwehrmechanismen*. Internationaler Psychoanalytischer Verlag, Wien, 1936.

S. Freud. The neuro-psychoses of defence. *Collected Papers*, 3:45–61, 1894. Originally published as: Die Abwehr-Neuropsychosen, 1894, Neurologisches Centralblatt, 13, 4, 50–51, 54–61.

S. Freud. The future prospects of psycho-analytic therapy. *Collected Papers*, 2:285–296, 1910. Originally published as: Über die zukünftigen Chancen der psychoanalytischen Therapie, 1910, Zentralblatt für Psychoanalyse, 1, 7, 297–311.

S. Freud. *Totem und Tabu: Einige Übereinstimmungen im Seelenleben der Wilden und der Neurotiker*. Hugo Heller & Cie, Leipzig und Wien, 1913.

V. Frey and A. Van de Rijt. Social influence undermines the wisdom of the crowd in sequential decision making. *Management science*, 67(7):4273–4286, 2021.

G. O. Gabbard and E. P. Lester. *Boundaries and boundary violations in psychoanalysis*. American Psychiatric Publishing, 1995.

I. Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

I. Gabriel, G. Keeling, A. Manzini, and J. Evans. We need a new ethics for a world of AI agents. *Nature*, 644(8075):38–40, Aug. 2025. doi: 10.1038/d41586--025-02454--5. URL `https://www.nature.com/articles/d41586--025-02454--5`. Comment.

H.-G. Gadamer. *Wahrheit und Methode*. J.C.B. Mohr (Paul Siebeck), 1960.

D. D. Garrett, G. R. Samanez-Larkin, S. W. MacDonald, U. Lindenberger, A. R. McIntosh, and C. L. Grady. The bold brain: greater variability of bold t2* signal is associated with better cognitive performance. *Journal of Neuroscience*, 33(2):835–840, 2013.

M. A. Garrett. Is artificial intelligence the great filter that makes advanced technical civilisations rare in the universe? *Acta Astronautica*, 219:731–735, 2024.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

M. Gell-Mann. *The Quark and the Jaguar: Adventures in the Simple and the Complex*. Macmillan, 1994.

G. Gigerenzer and W. Gaissmaier. Heuristic decision making. *Annual review of psychology*, 62: 451–482, 2011.

E. Goffman. *Frame analysis: An essay on the organization of experience*. Harvard university press, 1974.

B. Goldacre. *Bad pharma: how drug companies mislead doctors and harm patients*. Macmillan, 2014.

U. Goodenough and J. Heitman. Origins of eukaryotic sexual reproduction. *Cold Spring Harbor perspectives in biology*, 6(3):a016154, 2014.

I. Gough. Universal basic services: A theoretical and moral framework. *The Political Quarterly*, 90(3):534–542, 2019.

C. L. Grady and D. D. Garrett. Understanding variability in the bold signal and why it matters for aging. *Brain Imaging and Behavior*, 8:274–282, 2014. doi: 10.1007/s11682--013-9253--0.

J. Greene. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. The Penguin Press, New York, NY, 2013. ISBN 9781594202605.

M. Grey and C.-R. Segerie. Safety by measurement: a systematic literature review of ai safety evaluation methods. *arXiv preprint arXiv:2505.05541*, 2025.

S. J. Grossman and O. D. Hart. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of political economy*, 94(4):691–719, 1986.

K. Hackenburg, B. M. Tappin, L. Hewitt, E. Saunders, S. Black, H. Lin, C. Fist, H. Margetts, D. G. Rand, and C. Summerfield. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777):eaea3884, 2025.

D. Hadfield-Menell and G. K. Hadfield. Incomplete contracting and ai alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422, 2019.

D. Hadfield-Menell, A. D. Dragan, P. Fisac, and S. Russell. Cooperative inverse reinforcement learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 3909–3917, 2016.

D. Hadfield-Menell, A. D. Dragan, and S. Russell. The off-switch game. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI 2017)*, pages 220–227, 2017.

J. Haidt. *The righteous mind: Why good people are divided by politics and religion*. Vintage, 2012.

A. G. Haldane. Rethinking the financial network. In *Fragile stabilität–stabile fragilität*, pages 243–278. Springer, 2013.

C. Haldeman and J. M. Beggs. Critical branching captures activity in living neural networks and maximizes the number of metastable states. *Physical Review Letters*, 94(5):058101, 2005. doi: 10.1103/PhysRevLett.94.058101.

C. Hammond, H. Bergman, and P. Brown. Pathological synchronization in parkinson's disease: networks, models and treatments. *Trends in neurosciences*, 30(7):357–364, 2007.

F. M. Harold. *The way of the cell: molecules, organisms, and the order of life*. Oxford University Press, 2001.

H. K. Hartline, H. G. Wagner, and F. Ratliff. Inhibition in the eye of limulus. *The Journal of general physiology*, 39(5):651–673, 1956.

G. Hatano and K. Inagaki. Two courses of expertise. *Clinical Center for Early Childhood Development Annual Report*, 6:27–36, 1984.

K. E. Hauer, L. Edgar, S. O. Hogan, B. Kinnear, and E. Warm. The science of effective group process: lessons for clinical competency committees. *Journal of Graduate Medical Education*, 13(2 Suppl):59, 2021.

P. H. Hawley. Prosocial and coercive configurations of resource control in early adolescence: A case for the well-adapted machiavellian. *Merrill-Palmer Quarterly*, 49(3):279–309, 2003.

M. Heidegger. Letter on humanism. In W. McNeill, editor, *Pathmarks*. Cambridge University Press, Cambridge, 1998. Originally written 1946.

A. Heinz, G. K. Murray, F. Schlagenhauf, P. Sterzer, A. A. Grace, and J. A. Waltz. Towards a unifying cognitive, neurophysiological, and computational neuroscience account of schizophrenia. *Schizophrenia bulletin*, 45(5):1092–1100, 2019.

P. J. Hellyer, G. Scott, M. Shanahan, D. J. Sharp, and R. Leech. Cognitive flexibility through metastable neural dynamics is disrupted by damage to the structural connectome. *Journal of Neuroscience*, 35(24):9050–9063, 2015.

J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath. In search of homo economicus: behavioral experiments in 15 small-scale societies. *American economic review*, 91(2):73–78, 2001.

W. A. Hershberger. An approach through the looking-glass. *Animal Learning & Behavior*, 14 (4):443–451, 1986.

S. M. Herzog and R. Hertwig. Harnessing the wisdom of the inner crowd. *Trends in cognitive sciences*, 18(10):504–506, 2014.

W. G. Hill and A. Robertson. The effect of linkage on limits to artificial selection. *Genetical Research*, 8(3):269–294, 1966. PMID: 5980116.

T. T. Hills, P. M. Todd, D. Lazer, A. D. Redish, and I. D. Couzin. Exploration versus exploitation in space, mind, and society. *Trends in cognitive sciences*, 19(1):46–54, 2015.

D. R. Hofstadter. Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*, pages 499–538, 2001.

R. M. Hogarth. A note on aggregating opinions. *Organizational behavior and human performance*, 21(1):40–46, 1978.

J. H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. University of Michigan Press, Ann Arbor, MI, 1975. ISBN 0472084607.

C. S. Holling. Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics*, 4:1–23, 1973.

Honeywell. Honeywell and Google Cloud to accelerate autonomous operations with AI agents for the industrial sector, Oct. 2024. URL `https://www.honeywell.com/us/en/press/2024/10/honeywell-and-google-cloud-to-accelerate-autonomous-operations-with-ai-agents-for-the-` Press Release.

L. Hong and S. E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004.

D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

E. Hubinger, C. van Merwijk, V. Mikulik, J. Skreseki, and S. Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019. URL `https://arxiv.org/abs/1906.01820`.

R. R. Hudson and N. L. Kaplan. Deleterious background selection with recombination. *Genetics*, 141(4):1605–1617, 1995.

A. A. Hung and C. R. Plott. Information cascades: Replication and an extension to majority rule and conformity-rewarding institutions. *American Economic Review*, 91(5):1508–1520, 2001.

J. S. Isaacson and M. Scanziani. How inhibition shapes cortical activity. *Neuron*, 72(2):231–243, 2011.

D. Jablonski. Mass extinctions and macroevolution. *Paleobiology*, 31(S2):192–210, 2005.

I. L. Janis. *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes.* Houghton Mifflin, 1972.

K. Jaspers. *General psychopathology*, volume 2. JHU Press, 1997.

K. Javed and R. S. Sutton. The big world hypothesis and its ramifications for artificial intelligence. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*, 2024.

N. Jayanthi, C. Pinkham, L. Dugas, B. Patrick, and C. LaBella. Sports specialization in young athletes: evidence-based recommendations. *Sports health*, 5(3):251–257, 2013.

O. Jensen and A. Mazaheri. Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Frontiers in human neuroscience*, 4:186, 2010.

H. J. Jeon, S. Milli, and A. Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33:4415–4426, 2020.

Y. J. John, L. Caldwell, D. E. McCoy, and O. Braganza. Dead rats, dopamine, performance metrics, and peacock tails: Proxy failure is an inherent risk in goal-oriented systems. *Behavioral and Brain Sciences*, pages 1–68, 2023.

A. A. Kane, L. Argote, and J. M. Levine. Knowledge transfer between groups via personnel rotation: Effects of social identity and knowledge quality. *Organizational behavior and human decision processes*, 96(1):56–71, 2005.

T. B. Kashdan and J. Rottenberg. Psychological flexibility as a fundamental aspect of health. *Clinical psychology review*, 30(7):865–878, 2010.

S. A. Kauffman. The origins of order: Self-organization and selection in evolution. In *Spin glasses and biology*, pages 61–100. World Scientific, 1992.

P. D. Keightley and S. P. Otto. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature*, 443(7107):89–92, 2006.

S. Kerr. On the folly of rewarding a, while hoping for b. *Academy of Management journal*, 18 (4):769–783, 1975.

M. Kirchhoff, T. Parr, E. Palacios, K. Friston, and J. Kiverstein. The markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The royal society interface*, 15(138):20170792, 2018.

R. Kirk, I. Mediratta, C. Nalmpantis, J. Luketina, E. Hambro, E. Grefenstette, and R. Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*, 2023.

J. Kleinberg and M. Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

W. Klimesch, P. Sauseng, and S. Hanslmayr. Eeg alpha oscillations: the inhibition–timing hypothesis. *Brain research reviews*, 53(1):63–88, 2007.

E. Kolbert. *The sixth extinction: An unnatural history*. Henry Holt and Company, 2014.

A. Korzybski. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. The International Non-Aristotelian Library Publishing Company, Lancaster, PA, 1933.

N. Kosmyna, E. Hauptmann, Y. T. Yuan, J. Situ, X.-H. Liao, A. V. Beresnitzky, I. Braunstein, and P. Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 2025.

S. Kotler, M. Mannino, K. Friston, G. Buzsáki, J. S. Kelso, and G. Dumas. Pathfinding: a neurodynamical account of intuition. *Communications Biology*, 8(1):1214, 2025.

V. Krakovna, A. Gleave, and J. Miller. Specification gaming: The flip side of AI ingenuity. DeepMind Safety Research Blog, May 2020. URL `https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity`. Accessed on 2025–10-09.

S. W. Kraus, V. Voon, and M. N. Potenza. Should compulsive sexual behavior be considered an addiction? *Addiction*, 111(12):2097–2106, 2016.

T. S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 2nd edition, 1970. ISBN 9780226458083.

A. Kumar, J. Clune, J. Lehman, and K. O. Stanley. Questioning representational optimism in deep learning: The fractured entangled representation hypothesis. *arXiv preprint arXiv:2505.11581*, 2025.

Z. Kurth-Nelson, T. Behrens, G. Wayne, K. Miller, L. Luettgau, R. Dolan, Y. Liu, and P. Schwartenbeck. Replay and compositional computation. *Neuron*, 111(4):454–469, 2023.

Z. Kurth-Nelson, S. Sullivan, J. Z. Leibo, and M. Guitart-Masip. Dynamic diversity is the answer to proxy failure. *Behavioral and Brain Sciences*, 47:e77, 2024.

K. K. Ladha. The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, pages 617–634, 1992.

B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. doi: 10.1126/science.aab3050.

G. Lakoff and M. Johnson. *Metaphors We Live By*. University of Chicago Press, 1980.

N. Lane. *Vital question: energy, evolution, and the origins of complex life*. WW Norton & Company, 2015.

C. G. Langton. Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D: nonlinear phenomena*, 42(1-3):12–37, 1990.

S. Lazar and A. Nelson. Ai safety on whose terms?, 2023.

C. R. Leana. A partial test of janis' groupthink model: Effects of group cohesiveness and leader behavior on defective decision making. *Journal of management*, 11(1):5–18, 1985.

J. Lehtonen, M. D. Jennions, and H. Kokko. The many costs of sex. *Trends in ecology & evolution*, 27(3):172–178, 2012.

H. Lerner. *The dance of anger*. Harper & Row, 1985.

J. K. Leutgeb, S. Leutgeb, M.-B. Moser, and E. I. Moser. Pattern separation in the dentate gyrus and ca3 of the hippocampus. *science*, 315(5814):961–966, 2007.

C. Lévi-Strauss. *Les structures élémentaires de la parenté*. Presses Universitaires de France, Paris, 1949.

S. Lewandowsky, U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131, 2012.

L. A. Lipsitz and A. L. Goldberger. Loss of 'complexity' and aging: Potential applications of fractals and chaos theory to senescence. *JAMA*, 267(13):1806–1809, 1992. doi: 10.1001/jama.1992.03480130122036.

J. E. Lisman and O. Jensen. The theta-gamma neural code. *Neuron*, 77(6):1002–1016, 2013.

N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

A. Livnat, C. Papadimitriou, J. Dushoff, and M. W. Feldman. A mixability theory for the role of sex in evolution. *Proceedings of the National Academy of Sciences*, 105(50):19803–19808, 2008.

A. Livnat, C. Papadimitriou, N. Pippenger, and M. W. Feldman. Sex, mixability, and modularity. *Proceedings of the National Academy of Sciences*, 107(4):1452–1457, 2010.

M. C. Mackey and L. Glass. Oscillation and chaos in physiological control systems. *Science*, 197(4300):287–289, 1977. doi: 10.1126/science.267326.

L. Margulis and D. Sagan. *Microcosmos: Four Billion Years of Microbial Evolution*. Summit Books, New York, 1986.

L. Margulis and D. Sagan. *What Is Life?* Simon and Schuster, New York, 1995.

D. Marr. Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 262(841):23–81, 1971. doi: 10.1098/rstb.1971.0078.

A. H. Maslow. A theory of human motivation. *Psychological review*, 50(4):370, 1943.

W. A. Mason, A. Jones, and R. L. Goldstone. Propagation of innovations in networked groups. *Journal of Experimental Psychology: General*, 137(3):422, 2008.

J. Maynard Smith. The origin and maintenance of sex. In *Group selection*, pages 163–175. Aldine Atherton, 1971.

J. Maynard Smith. *The evolution of sex*, volume 4. Cambridge University Press Cambridge, 1978.

M. McCain, R. Linthicum, C. Lubinski, A. Tamkin, S. Huang, M. Stern, K. Handa, E. Durmus, T. Neylon, S. Ritchie, K. Jagadish, P. Maheshwary, S. Heck, A. Sanderford, and D. Ganguli. How people use Claude for support, advice, and companionship. Anthropic, June 2025. URL `https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship`.

J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

B. L. McNaughton and R. G. M. Morris. Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences*, 10(10):408–415, 1987. doi: 10.1016/0166--2236(87)90011--8.

R. K. Merton. Bureaucratic structure and personality. *Social Forces*, 18(4):560–568, 1940. doi: 10.2307/2570634.

L. Messeri and M. J. Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, 2024.

E. K. Miller and J. D. Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.

G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956. doi: 10.1037/h0043158.

G. A. Miller, G. Eugene, and K. H. Pribram. *Plans and the Structure of Behaviour*. Routledge, 1960.

S. Milli, D. Hadfield-Menell, A. Dragan, and S. Russell. Should robots be obedient? *arXiv preprint arXiv:1705.09990*, 2017.

S. Minuchin. *Families and Family Therapy*. Harvard University Press, 1974.

A. L. Mishara. Klaus conrad (1905–1961): Delusional mood, psychosis, and beginning schizophrenia. *Schizophrenia Bulletin*, 36(1):9–13, 2010.

A. Miyake, N. P. Friedman, M. J. Emerson, A. H. Witzki, A. Howerter, and T. D. Wager. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive psychology*, 41(1):49–100, 2000.

H. J. Muller. Some genetic aspects of sex. *The American Naturalist*, 66(703):118–138, 1932.

P. M. Müller, G. Miron, M. Holtkamp, and C. Meisel. Critical dynamics predicts cognitive performance and provides a common framework for heterogeneous mechanisms impacting cognition. *Proceedings of the National Academy of Sciences*, 122(14):e2417117122, 2025.

R. U. Muller and J. L. Kubie. The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *The Journal of Neuroscience*, 7(7):1951–1968, 1987. doi: 10.1523/JNEUROSCI.07--07-01951.1987.

I. Murdoch and M. Midgley. *The sovereignty of good*. Routledge, 2013.

L. Myllyaho, M. Raatikainen, T. Männistö, T. Mikkonen, and J. K. Nurminen. Systematic literature review of validation methods for ai systems. *arXiv preprint arXiv:2107.12190*, 2021.

F. Nietzsche. *Also sprach Zarathustra: Ein Buch für Alle und Keinen*. Ernst Schmeitzner, Chemnitz, 1883. Published in four parts. Parts 1–3 (1883–1884) by Schmeitzner; Part 4 (1885) privately printed by the author.

M. C. Nussbaum. *The fragility of goodness: Luck and ethics in Greek tragedy and philosophy*. Cambridge University Press, 2001.

J. P. O'Doherty, J. Cockburn, and W. M. Pauli. Learning, reward, and decision making. *Annual review of psychology*, 68(1):73–100, 2017.

S. Ohlsson. Information-processing explanations of insight and related phenomena. *Advances in the psychology of thinking*, 1:1–44, 1992.

C. O'Keefe, P. Cihon, B. Garfinkel, C. Flynn, J. Leung, and A. Dafoe. The windfall clause: Distributing the benefits of ai for the common good. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 327–331, 2020.

C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

M. Olson. *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities*. Yale University Press, New Haven, 1982.

OpenAI. Democratic inputs to ai. `https://openai.com/blog/democratic-inputs-to-ai/`, May 2023. Accessed: 2025-12-05.

OpenAI. The state of enterprise AI 2025 report. Technical report, OpenAI, 2025. URL `https://cdn.openai.com/pdf/7ef17d82-96bf-4dd1-9df2-228f7f377a29/the-state-of-enterprise-ai_2025-report.pdf`.

L. D. Ordóñez, M. E. Schweitzer, A. D. Galinsky, and M. H. Bazerman. Goals gone wild: The systematic side effects of overprescribing goal setting. *Academy of Management Perspectives*, 23(1):6–16, 2009.

R. C. O'Reilly, T. E. Hazy, J. Mollick, P. Mackie, and S. Herd. Goal-driven cognition in the brain: a computational framework. *arXiv preprint arXiv:1404.7591*, 2014.

M. Owen. How to avoid the problem of 'group-think' in your boardroom, December 2019. URL `https://owenmorrispartnership.com/how-to-avoid-the-problem-of-group-think-in-your-boardroom/`.

V. Padmakumar and H. He. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*, 2023.

F. Perls, R. Hefferline, and P. Goodman. *Gestalt Therapy: Excitement and Growth in the Human Personality*. Julian Press, 1951.

D. Pimentel, R. Zuniga, and D. Morrison. Update on the environmental and economic costs associated with alien-invasive species in the united states. *Ecological economics*, 52(3):273–288, 2005.

S. Pinker. *The Language Instinct: How the Mind Creates Language*. William Morrow and Company, New York, 1994.

M. Polanyi. *The Tacit Dimension*. Doubleday, Garden City, NY, 1966.

E. Polster and M. Polster. *Gestalt therapy integrated: Contours of theory & practice*, volume 6. Vintage, 1974.

K. R. Popper. *Logik der Forschung: Zur Erkenntnistheorie der modernen Naturwissenschaft*. Verlag von Julius Springer, Wien (Vienna), 1934.

I. Prigogine and I. Stengers. *Order Out of Chaos: Man's New Dialogue with Nature*. Bantam Books, New York, 1984.

M. I. Rabinovich, R. Huerta, P. Varona, and V. S. Afraimovich. Transient cognitive dynamics, metastability, and decision making. *PLoS Computational Biology*, 4(5):e1000072, 2008. doi: 10.1371/journal.pcbi.1000072.

D. M. Raup. The role of extinction in evolution. *Proceedings of the National Academy of Sciences*, 91(15):6758–6763, 1994.

W. Reich. On character analysis. *The Psychoanalytic Review (1913-1957)*, 20:89, 1933.

L. Reynolds and K. McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7, 2021.

W. J. Ripple and R. L. Beschta. Trophic cascades in yellowstone: the first 15 years after wolf reintroduction. *Biological Conservation*, 145(1):205–213, 2012.

J. Rockström, W. Steffen, K. Noone, Å. Persson, F. S. Chapin III, E. F. Lambin, T. M. Lenton, M. Scheffer, C. Folke, H. J. Schellnhuber, et al. A safe operating space for humanity. *nature*, 461(7263):472–475, 2009.

F. Roux and P. J. Uhlhaas. Working memory and neural oscillations: alpha–gamma versus theta–gamma codes for distinct wm information? *Trends in cognitive sciences*, 18(1):16–25, 2014.

S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Publishing Group, 2019. ISBN 9780525558620. URL `https://books.google.co.uk/books?id=M1eFDwAAQBAJ`.

P. Saffo. Strong opinions weakly held, 2008. URL `https://saffo.com/02008/07/26/strong-opinions-weakly-held/`.

C. B. Saper and B. B. Lowell. The hypothalamus. *Current Biology*, 24(23):R1111–R1116, 2014.

B. T. Saunders and T. E. Robinson. The role of dopamine in the accumbens core in the expression of pavlovian-conditioned responses. *European Journal of Neuroscience*, 36(4): 2521–2532, 2012.

R. C. Schank and R. P. Abelson. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press, 1977.

P. Scharre. *Army of None: Autonomous Weapons and the Future of War*. W. W. Norton, New York, 2018.

E. Schrödinger. *What is Life? The Physical Aspect of the Living Cell*. Cambridge University Press, Cambridge, UK, 1944. Based on lectures delivered at Trinity College, Dublin, February 1943.

J. Schulkin and P. Sterling. Allostasis: a brain-centered, predictive mode of physiological regulation. *Trends in neurosciences*, 42(10):740–752, 2019.

J. C. Scott. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, CT, 1998. ISBN 9780300070163.

A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.

A. N. Sell. The recalibrational theory and violent anger. *Aggression and violent behavior*, 16 (5):381–389, 2011.

N. Selwyn. On the limits of artificial intelligence (ai) in education. *Nordisk tidsskrift for pedagogikk og kritikk*, 10(1):3–14, 2024.

T. V. Sewards and M. A. Sewards. Representations of motivational drives in mesial cortex, medial thalamus, hypothalamus and midbrain. *Brain research bulletin*, 61(1):25–49, 2003.

J. Y. Shah, R. Friedman, and A. W. Kruglanski. Forgetting all else: on the antecedents and consequences of goal shielding. *Journal of personality and social psychology*, 83(6):1261, 2002.

R. Shah, P. Freire, N. Alex, R. Freedman, D. Krasheninnikov, L. Chan, M. D. Dennis, P. Abbeel, A. Dragan, and S. Russell. Benefits of assistance over reward learning. *NeurIPS*, 2020.

D. Shapiro. *Neurotic styles*. Basic Books, 1965.

M. Sharp, O. Bilgin, I. Gabriel, and L. Hammond. Agentic inequality. *arXiv preprint arXiv:2510.16853*, 2025.

W. L. Shew, H. Yang, T. Petermann, R. Roy, and D. Plenz. Neuronal avalanches imply maximum dynamic range in cortical networks at criticality. *Journal of neuroscience*, 29(49): 15595–15600, 2009.

W. L. Shew, H. Yang, S. Yu, R. Roy, and D. Plenz. Information capacity and transmission are maximized in balanced cortical networks with neuronal avalanches. *Journal of Neuroscience*, 31(1):55–63, 2011. doi: 10.1523/JNEUROSCI.4637--10.2011.

V. Shiva. *Monocultures of the mind: Perspectives on biodiversity and biotechnology*. Palgrave Macmillan, 1993.

J. Sidanius and F. Pratto. *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press, 2001.

A. M. Sillito. The contribution of inhibitory mechanisms to the receptive field properties of neurones in the striate cortex of the cat. *The Journal of Physiology*, 250(2):305–329, 1975.

G. G. Simpson. *Tempo and Mode in Evolution*. Number 15 in Columbia Biological Series. Columbia University Press, New York, 1944.

P. Singer. *The expanding circle*. Clarendon Press Oxford, 1981.

A. Singla, A. Sukharevsky, L. Yee, M. Chui, B. Hall, and T. Balakrishnan. The state of AI in 2025: Agents, innovation, and transformation. Technical report, McKinsey & Company, Nov. 2025. URL `https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai`.

M. Sloane, E. Moss, O. Awomolo, and L. Forlano. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–6, 2022.

A. Smith. *An inquiry into the nature and causes of the wealth of nations: Volume One*. London: printed for W. Strahan; and T. Cadell, 1776., 1776.

M. J. Smith. *When I say no, I feel guilty*. Bantam, 1985.

P. Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1–2):159–216, 1990. doi: 10.1016/0004--3702(90)90007-M.

N. Soares and B. Fallenstein. Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report*, 8, 2014.

S. Sontag, C. Drew, and A. L. Drew. *Blind man's bluff: The untold story of American submarine espionage*. Public Affairs, 1998.

D. Speijer, J. Lukeš, and M. Eliáš. Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. *Proceedings of the National Academy of Sciences*, 112(29):8827–8834, 2015.

D. Sperber, F. Clément, C. Heintz, O. Mascaro, H. Mercier, G. Origgi, and D. Wilson. Epistemic vigilance. *Mind & language*, 25(4):359–393, 2010.

K. O. Stanley and J. Lehman. *Why Greatness Cannot Be Planned: The Myth of the Objective*. Springer, Cham, 2015. ISBN 978-3-319-15523-4.

K. O. Stanley, J. Lehman, and L. B. Soros. Open-endedness: The last grand challenge you've never heard of. O'Reilly Radar, 2017. URL `https://www.oreilly.com/radar/open-endedness-the-last-grand-challenge-youve-never-heard-of/`. Accessed: 2025-01-28.

G. Stasser and W. Titus. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of personality and social psychology*, 48(6):1467, 1985.

R. Stein. *Incest and human love: The betrayal of the soul in psychotherapy*. Third Press, 1973.

P. Sterling and J. Eyer. Allostasis: A new paradigm to explain arousal pathology. In S. Fisher and J. Reason, editors, *Handbook of Life Stress, Cognition and Health*, pages 629–649. John Wiley & Sons, Chichester, 1988.

S. G. Straus, A. M. Parker, and J. B. Bruce. The group matters: A review of processes and outcomes in intelligence analysis. *Group Dynamics: Theory, Research, and Practice*, 15(2): 128, 2011.

J. Surowiecki. *The wisdom of crowds*. Vintage, 2005.

D. Susskind. *A World Without Work: Technology, Automation, and How We Should Respond*. Metropolitan Books, New York, 2020.

R. I. Sutton and M. R. Louis. How selecting and socializing newcomers influences insiders. *Human Resource Management*, 26(3):347–361, 1987.

V. Tausk. Über die entstehung des 'beeinflussungsapparates' in der schizophrenie. *Internationale Zeitschrift für Psychoanalyse*, 5:1–33, 1919.

TechCrunch. Sam altman says ChatGPT has hit 800m weekly active users, Oct. 2025. URL `https://techcrunch.com/2025/10/06/sam-altman-says-chatgpt-has-hit-800m-weekly-active-users/`. Accessed: 2025-11-19.

B. J. Tepper. Consequences of abusive supervision. *Academy of management journal*, 43(2): 178–190, 2000.

P. E. Tetlock. A value pluralism model of ideological reasoning. *Journal of personality and social psychology*, 50(4):819, 1986.

D. Tilman. Biodiversity: population versus ecosystem stability. *Ecology*, 77(2):350–363, 1996.

E. Tognoli and J. S. Kelso. The metastable brain. *Neuron*, 81(1):35–48, 2014.

J. Tooby and L. Cosmides. The psychological foundations of culture. In J. H. Barkow, L. Cosmides, and J. Tooby, editors, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, pages 19–136. Oxford University Press, New York, NY, 1992.

D. S. Touretzky and G. E. Hinton. A distributed connectionist production system. *Cognitive Science*, 12(3):423–466, 1988. doi: 10.1207/s15516709cog1203_3.

H. M. Trainer, J. M. Jones, J. G. Pendergraft, C. K. Maupin, and D. R. Carter. Team membership change "events": A review and reconceptualization. *Group & Organization Management*, 45(2):219–251, 2020.

A. Treves and E. T. Rolls. Computational analysis of the role of the hippocampus in memory. *Hippocampus*, 4(3):374–391, 1994. doi: 10.1002/hipo.450040319.

A. M. Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37–72, 1952. doi: 10.1098/rstb.1952.0012.

N. Vafeas. Length of board tenure and outside director independence. *Journal of Business Finance & Accounting*, 30(7-8):1043–1064, 2003.

R. R. Vallacher and D. M. Wegner. What do people think they're doing? action identification and human behavior. *Psychological review*, 94(1):3, 1987.

M. Van Der Meer, Z. Kurth-Nelson, and A. D. Redish. Information processing in decision-making systems. *The Neuroscientist*, 18(4):342–359, 2012.

B. Van Valkenburgh, X. Wang, and J. Damuth. Cope's rule, hypercarnivory, and extinction in north american canids. *Science*, 306(5693):101–104, 2004.

F. J. Varela, E. Thompson, and E. Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, Cambridge, MA, 1991.

S. L. Videbeck. *Psychiatric-mental health nursing*. Lippincott Williams & Wilkins, 2010.

G. P. Wagner and L. Altenberg. Perspective: complex adaptations and the evolution of evolvability. *Evolution*, 50(3):967–976, 1996.

M. P. Warren, J. B. Gunn, L. H. Hamilton, L. F. Warren, and W. G. Hamilton. Scoliosis and fractures in young ballet dancers. *New England Journal of Medicine*, 314(21):1348–1353, 1986.

L. Waschke, N. A. Kloosterman, J. Obleser, and D. D. Garrett. Behavior needs neural variability. *Neuron*, 109(5):751–766, 2021.

H. Watson. Biological membranes. *Essays in biochemistry*, 59:43–69, 2015.

M. Weber. Die protestantische ethik und der geist des kapitalismus. *Archiv für Sozialwissenschaft und Sozialpolitik*, 20–21:1–54, 1–110, 1905.

A. Weismann. *Essays upon heredity and kindred biological problems*. Clarendon Press, Oxford, 1889.

P. Welinder, S. Branson, P. Perona, and S. Belongie. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23, 2010.

N. Wiener. Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science*, 131(3410): 1355–1358, 1960.

B. Williams. *Ethics and the Limits of Philosophy*. Harvard University Press, Cambridge, MA, 1985.

G. C. Williams. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton University Press, Princeton, NJ, 1966.

L. Wittgenstein. *Tractatus Logico-Philosophicus*. Routledge & Kegan Paul, London, 1922.

M. J. Wood, K. M. Douglas, and R. M. Sutton. Dead and alive: Beliefs in contradictory conspiracy theories. *Social psychological and personality science*, 3(6):767–773, 2012.

S. C. Woolley and P. N. Howard. *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press, 2018.

S. Wu, B. A. Nijstad, and Y. Yuan. Membership change, idea generation, and group creativity: A motivated information processing perspective. *Group Processes & Intergroup Relations*, 25 (5):1412–1434, 2022.

T. Wu. *The master switch: The rise and fall of information empires.* Vintage, 2011.

R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, and W. Xu. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*, 2024.

W. Xu, N. Jojic, S. Rao, C. Brockett, and B. Dolan. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *Proceedings of the National Academy of Sciences*, 122(35):e2504966122, 2025.

S. Yachi and M. Loreau. Biodiversity and ecosystem productivity in a fluctuating environment: the insurance hypothesis. *Proceedings of the National Academy of Sciences*, 96(4):1463–1468, 1999.

R. V. Yampolskiy. On monitorability of ai. *AI and Ethics*, 5(1):689–707, 2025.

G. M. Yontef. *Awareness, dialogue & process: Essays on Gestalt therapy.* The Gestalt Journal Press, 1993.

E. Yudkowsky. Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*, 2004.

Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021.

E. Zhou and D. Lee. Generative artificial intelligence, human creativity, and art. *PNAS nexus*, 3(3):pgae052, 2024.

S. Zhuang and D. Hadfield-Menell. Consequences of misaligned ai. *Advances in Neural Information Processing Systems*, 33:15763–15773, 2020.