What is Lock-In?

Alfie Lamerton

Epistemic status: a combination and synthesis of others' work, analysed and written over a few weeks. Early working definition that is open to criticism.

TL;DR

I create a definition of lock-in for use in future discussion and writing, and operationalise lock-in for future research. I define lock-in risks as *the probabilities of situations in which features of the world, typically negative elements of human culture, are made stable for long periods of time.*

Why Define Lock-In?

Lock-in is the central theme of Formation Research. Therefore before conducting any research on lock-in, it is important to create a strong working definition for use in subsequent discussion. The stronger the definition, the stronger the foundation on which the subsequent work will rest, and the more specific and explicit we can be when tackling lock-in.

Key Thinkers

Nick Bostrom

Bostrom has introduced a number of relevant concepts to discussions about lock-in. He introduces the notion of a Singleton – a world order in which there is a single decision-making at the highest level (Bostrom, 2005). He also introduces the idea of a decisive strategic advantage, a situation in which one entity may gain strategic power over the fate of humanity at large. He relates this to the potential formation of a Singleton (Bostrom, 2014).

He also introduces the instrumental convergence hypothesis, providing insight into potential motivations of autonomous AI systems. The hypothesis suggests a number of logically implied goals an agent will develop when given an initial goal. Lastly, he introduces the value loading problem, and the risks of misalignment due to issues such as goal misspecification.

William MacAskill

MacAskill talks about lock-in in a chapter of What We Owe the Future (MacAskill, 2022). He uses the Hundred Schools of Thought period in ancient China, the concept of writing, and the potential for artificial general intelligence to define value lock-in and argue for its potential. He defines lock-in as a situation where a single ideology or set of ideologies, or value system takes control of an area of the world, or the whole world, and persists for an extremely long time.

Toby Ord

In his chapter on future risks in the Precipice (Ord, 2020), Ord categorises potential existential catastrophes for humanity into extinction, and failed continuation. He then categorises failed continuation into the unrecoverable collapse of civilisation, and unrecoverable dystopia.

He then divides the unrecoverable dystopias into three categories: undesired dystopia, enforced dystopia, and desired dystopia.

Undesired dystopia refers to a situation where 'people don't want that world, yet the structure of society makes it almost impossible for them to coordinate to change it'. Enforced dystopia is where 'only a small group wants that world but enforces it against the wishes of the rest'. Desired dystopia refers to a situation where people 'do want that world, yet they are misguided and the world falls far short of what they could have achieved.'

Instead of defining these situations as lock-ins, Ord uses these potential futures as categories of *existential catastrophes*, which are a distinct concept with some overlap with the notion of lock-in. He defines these regimes as being a negative turning point in the history of human potential that increases the likelihood of a 'worthy future'.

Lukas Finnveden

In an interview with Future Matters, Lukas Finnveden, main author on AGI and Lock-In (Finnveden et al., 2022) and research analyst at Open Philanthropy, defines lock-in in terms of predictable stability (Stafforini & Vandermerwe, 2023). 'Some property of the world is locked in if it's very probable that the property of the world will hold true for a very large amount of time'.

Jess Riedel

Jess Riedel, co-author of AGI and lock-in, physicist at NTT Research, uses the term *singleton* introduced by Nick Bostrom in What is a Singleton? (Bostrom, 2005) to create a definition of lock-in (Riedel, 2021). If there is an AGI singleton (world order in which there is a single decision-making agency at the highest level) with stable values (its values do not change over time), then those values are locked-in. This can be further defined as global lock-in if it applies to all earth-originating life.

Carl Shulman, co-author on AGI and lock-in, has also done a lot of thinking on lock-in, but does not provide a concise definition in his public discussions.

Analogies and Related Concepts

Technology Lock-In

Ways in which technologies can get locked in by markets and evolutionary forces, leading to stability that is not necessarily optimal. For example, the QWERTY keyboard is commonplace despite being a suboptimal layout for typing in English.

Moloch

Symbol of systems that demand constant sacrifice, causing the depletion of human potential, and analogy for value lock-in in that society can be locked into values which require continuous sacrifice e.g., resource exploitation requiring constant environmental degradation. Scott Alexander lists 10 real-world examples of *multipolar traps* in Meditations on Moloch to illustrate this analogy (Alexander, 2014).

Static Society

A society characterised by minimal change or evolution over long periods. In The Beginning of Infinity, David Deutsche describes societies as static (values held in place with little variation for long periods) and dynamic (inversely, values updating in the direction of what we call progress over short periods with no defined endpoint) (Deutsch, 2012).

The Malthusian Trap

Self-reinforcing dynamics that can be likened to lock-in dynamics. Resource depletion limits growth, and constrains progress and improvement without paradigm-shifting innovation.

Concepts from economics and game theory such as vendor lock-in, tragedy of the commons, path dependence, equilibrium lock-in, repeated games, and deadlocks, also provide models for thinking about how situations can get locked in.

The End of History

In political philosophy, the end of history refers to a concept presented by Francis Fukuyama in the late 20th century in his book “The End of History and the Last Man”. The idea points at the endpoint of humanity’s sociocultural evolution due to reaching its optimal configuration. It is possible that in the future, the evolution of culture will reach a point where elements of human culture will no longer change.

Steady State

In systems theory, a steady state defines the behaviour of a system when its key variables remain constant over time.

Definition

For our purposes, we define lock-in as a *situation where a feature of the world is stable for a long time.* Our definition of lock-in follows from, and as is most aligned with, Lukas Finnveden's definition because that definition targets the long-term and neutral nature of what we consider to be a lock-in, and we believe this is the best explanation for what a lock-in is.

Operationalising Lock-In

Just as the definition is necessary for discussing lock-in qualitatively, so operationalising lock-in is necessary for researching lock-in quantitatively. We start by transforming our qualitative definition of lock-in into an objective physical phenomenon. We say a lock-in is a situation where a feature of the world is made stable, within reasonable error bounds, for a long period of time.

For example, under this definition, the Great Wall of China can be described as a lock-in, due to its stability over 2,600 years. Another example like this is the Pyramids of Giza, which have been held physically stable for ~4,500 years.

We can say feature x is locked in if it remains constant in the interval [t1, t2] where t2 » t1. A world in which x exists can be described as a lock-in. So in researching lock-in risks, we are focusing on the probability of some world in which an x *that we care about* manifests.

To decide if a feature of the world has changed or not, idealistically we would make x a constant and say any change means it is not stable. But for measuring the world, where stability is fuzzier, we introduce error bounds.

To illustrate and justify the inclusion of error boundaries on stability, note that probably not every brick in the Great Wall of China or every grain of sand in the Pyramids of Giza are in the same place as they were when first placed there. Similarly, it may not be the case that every explicit value remains identical over the course of a future totalitarian regime.

Nonetheless, within *reasonable* error boundaries, we say these features of the world can still be described as locked-in. We still call the wall the wall, and the pyramids the pyramids, even if 0.1% of the elements they were composed of have moved or changed.[1]

[1] At the limit, this could drift into the Ship of Theseus Paradox, but the error-boundaries will probably never be large enough.

We define the error boundaries for lock-in as the range of situations where we would describe the feature as still being fundamentally the same. In studies we will set the error bounds on a case-by-case basis, as it does not seem clear that one value will apply to all studies of lock-in. Therefore, if any independent measurement of x falls within these bounds, we say it is stable.

To ensure lock-in is always measurable, it follows that lock-in cannot refer to any phenomenon that cannot be empirically measured. For example, qualia and subjective experience are hard to measure. However, proxy measures such as self-reports/experience sampling make even these concepts at least somewhat tractable. When dealing with conceptual attributes such as values, it is important to find a measurable variable that represents these values, otherwise we cannot empirically study lock-in.

This lets us study lock-in quantitatively in the physical world and in computational simulations. But that doesn't answer the question of which lock-in we care about. The persistence of the Great Wall of China and Pyramids of Giza are probably not relevant to the long-term future of humanity. The persistence of other features of the world, however, could well be. We argue for which features of the world those are.

Features of the World We Care About

The things we care most about when wanting to minimise lock-in risks (from our point of view as humans) are not all the features of the world, but the features of *human culture*, namely values and ethics, power and decision-making structures, cultural norms and ideologies, and scientific and technological progress. From the perspective of minimising risk, we are therefore most interested in potential lock-in scenarios that would be *harmful, oppressive, persistent (long-term or final), or widespread.*

Positive, Negative, and Neutral Lock-In

So now the definition looks something like *bad situations for our culture that last a long time.* But what about Ord's focus on desirability, and Jess Riedel's focus on AGI? Our definition now attempts to unify these attributes by categorising lock-ins. Let's start by distinguishing between positive, negative, and neutral lock-ins.

Positive Lock-In

Positive lock-ins are tricky, because as MacAskill points out, today's positive lock-in might be considered parochial and undesirable by future generations. Just as locking in the values of slavery would be seen as a terrible idea today, so locking in some values of today might be seen as a terrible idea in the future. This is persistently a paradox, because in a society that makes constant progress towards an improved version of itself, there may never be a point at which we can be comfortable locking in anything.

However, there is a small space of potential solutions where we might be able to converge on something close to positive lock-ins. This is the region where we lock-in features of human culture that we believe contribute to the minimisation of lock-in risks.

One example is human extinction. Efforts to prevent this from happening, such as those in the field of existential risk, can be argued as being good values, because they prevent us from a persistent situation – all being dead. Another example is locking in the value of *never banning criticism*, because this prevents us from getting into a situation where critical thinking cannot be targeted at a stable feature of the world to question its integrity or authority. Being able to criticise anything means there is always room for a stable regime to be overhauled. Lastly, the *preservation of sustainable competition* could be argued as a positive lock-in, because it helps prevent the monopolisation of features of culture, such as in markets.

The concept of a positive lock-in is delicate, and further work is needed to learn whether we can sensibly converge on positive lock-ins to mitigate risks posed by other lock-ins.

Neutral Lock-In

We define these as lock-ins that we are not particularly interested in. As mentioned, the openness of our definition allows for many features of the world to be considered lock-ins. For example, the temperature of Earth, or more specific to human culture, the concept of work. These are features of the work which tend to remain stable, but that we are not trying to make more or less stable.

Negative Lock-In

These are the lock-ins we are most interested in. The formation of these kinds of lock-ins would have negative implications for humanity. As mentioned, we care most about lock-ins that are:

1. **Harmful**: resulting in physical or psychological harm to individuals
2. **Oppressive**: suppressing individuals' freedom, autonomy, speech, or opportunities, or the continued evolution of culture
3. **Persistent**: long-term, unrecoverable, or irreversible
4. **Widespread**: concerning a significant portion of individuals relative to the total population

While there are other negative qualities of potential lock-ins, these are the qualities we care most about. Some examples of other areas we might care about are situations where humanity is limited in happiness, freedom, rights, well-being, quality of life, meaning, autonomy, survival, or empowerment.

Human-Only, AI-Enabled, and AI-Led Lock-In

So we have established a categorisation of lock-in in terms of the desirability of outcomes for individuals. Now let's factor in AI. As most key thinkers identify, AI, and especially AGI, plays an important role in hypothetical future lock-ins. There are many reasons for this, such as improved surveillance, error correction, human longevity and immortality, and power-seeking behaviour and instrumental convergence.

It is important to address this fact by further categorising lock-ins by their relationship to AI systems. We created three such categorisations:

1. **Human-only**: led by a human or group of humans and not enhanced significantly by an AI system
2. **AI-enabled**: led by a human or a group of humans leveraging an AI system
3. **AI-led**: led by an AI system or group of AI systems

Human-only lock-ins have existed in the past. The best examples of past human-only lock-ins are the expansionist totalitarian regimes led by Adolf Hitler, Joseph Stalin, or Benito Mussolini.

There are no concrete existing examples of AI-enabled lock-ins yet, but some related examples are the Molochian forces of recommendation algorithms and the advent of short-form video on human attention and preferences by companies such as Meta and TikTok, in products such as the TikTok app, Instagram Reels, and YouTube shorts.

These are situations where intelligent algorithms have been optimised by humans for customer retention. There are also no examples of AI-led lock-ins yet, but AI takeover and misalignment scenarios have been considered by the AI alignment community.

Accidental and Deliberate Lock-Ins

One last way these categories can be further refined is by considering them accidental or deliberate. An accidental lock-in would be a lock-in that emerges without intention by humans, while a deliberate lock-in would be one which emerges as a consequence of human intention to do so.

So lock-in risks are probabilities of negative lock-ins manifesting in the world. Thus, minimising lock-in risks means to reduce these probabilities; or, *minimise the likelihood that elements of culture that are human-only, AI-enabled, or AI-led, become stable to the extent that they are harmful, oppressive, persistent, or widespread, either accidentally or deliberately.*

Conclusion

This document surveys the existing literature and thinking on lock-in, highlighting the contributions made by key thinkers in the area, and adjacent concepts. A definition of lock-in is then synthesised in light of the existing definitions, and categories are defined to make different lock-ins explicit in terms of scope, desirability, and AI. This document aims to be the starting point for further work on lock-in risks by creating a foundational definition of lock-in that such work can use.

References

1. Alexander, S. (2014). *Meditations On Moloch | Slate Star Codex.* https://slatestarcodex.com/2014/07/30/meditations-on-moloch/
2. Bostrom, N. (2005). *What is a Singleton?* What Is a Singleton? https://nickbostrom.com/fut/singleton
3. Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press. https://books.google.co.uk/books?id=7_H8AwAAQBAJ
4. Deutsch, D. (2012). *The Beginning of Infinity: Explanations that Transform the World.* Penguin UK.
5. Finnveden, L., Jess*Riedel, & CarlShulman. (2022). _AGI and Lock-In*. https://forum.effectivealtruism.org/posts/KqCybin8rtfP3qztq/agi-and-lock-in
6. MacAskill, W. (2022). *What We Owe the Future.* Basic Books. https://books.google.co.uk/books?id=nd_GzgEACAAJ
7. Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity.* Hachette Books. https://books.google.co.uk/books?id=3aSiDwAAQBAJ
8. Riedel, J. (2021). *Value Lock-in Notes 2021.*
9. Stafforini, P., & Vandermerwe, M. (2023). *Future Matters #7: AI timelines, AI skepticism, and lock-in.* https://forum.effectivealtruism.org/posts/Ky7C7whxdLexXWqss/future-matters-7-ai-timelines-ai-skepticism-and-lock-in