An Axiomatic Version of Fitch's Paradox

Abstract. A variation of Fitch's Paradox is given, where no special rules of inference are assumed, only axioms. These axioms follow from the familiar assumptions which involve rules of inference. We show (by constructing a model) that by allowing that possibly the knower doesn't know his own soundness (while still requiring he be sound), Fitch's Paradox is avoided. Provided one is willing to admit that sound knowers may be ignorant of their own soundness, this might offer a way out of the paradox.

1. Introduction

Fitch's Paradox is the fact that certain sets of assumptions imply, by what is called the Church-Fitch argument, that $\phi \rightarrow K(\phi)$. There are various different choices for the sets of assumptions, and we will initially start from the following one, call it F:

- T: $K(\phi) \rightarrow \phi$
- $\wedge$: $K(\phi \wedge \psi) \rightarrow K(\phi) \wedge K(\psi)$
- C: $\phi \rightarrow P(K(\phi))$.
- D-Rule: From $\neg\phi$, infer $\neg P(\phi)$.

It is common to replace D-Rule with the axiom $K(\neg\phi) \rightarrow \neg P(\phi)$ while adding the rule of necessitation $\phi / K(\phi)$ (this is essentially what Salerno (2010) does), but F is more basic, being easily implied by the alternative. Following Chow (1998), we may say the following about the paradox (and about paradoxes in general): to *resolve* it is to weaken or alter the assumptions so that the unreasonable conclusion no longer follows. Therefore there may be many different resolutions to a paradox, and it may be subjective which is better, if any. Accordingly, in this paper I will demonstrate one resolution, and, though I will argue for its philosophical plausibility, I do not intend to champion it above other known resolutions.

The assumption I will target for removal is $K(K(\phi) \rightarrow \phi)$. The reader may object that this is not one of the assumptions listed! It merely follows from them. That it follows from them is a special case of the Church-Fitch argument, if nothing else. Which leads us to ask:

Informal Question: Can $K(K(\phi) \rightarrow \phi)$ be proved from F by a different method than variation on the Church-Fitch argument?

I have not managed to do so, and I very tentatively conjecture the answer is "no", though I do not know for sure. Suppose that the answer really is "no". Then any resolution which blocks the Church-Fitch argument and goes no further, will necessarily block all the proofs of $K(K(\phi) \rightarrow \phi)$, unless there be something special about this particular instance of the argument, against the spirit of the Question. If the Answer is indeed "no", then an optimal resolution of the paradox *must* eliminate $K(K(\phi) \rightarrow \phi)$ as a consequence of F (or of whatever system replaces F). For example, from F, the intuitionist can deduce $\neg\neg K(K(\phi) \rightarrow \phi)$, following Church and Fitch til the end. To cover the final gap, the intuitionist needs a whole new tactic, whose existence would positively answer the Question. Maybe the intuitionist could simply add $K(K(\phi) \rightarrow \phi)$ as an axiom in addition to letting intuitionism destroy the paradox, but that resolution would not be "optimal". The point of all this is to justify the choice of targeting $K(K(\phi) \rightarrow \phi)$: if it has to go in any case, why not aim at it directly?

My aim is to break up the rules of inference, which can be thought of as huge indivisible assumptions, into smaller pieces (together capable of taking the rules' place). More smaller assumptions means more flexibility choosing how to resolve the paradox, unfortunately such new resolutions will probably not translate into resolutions for the familiar form of the paradox (we are saddened that this may justifiably render our result an outlier in the bigger picture of Fitch's Paradox). The smaller assumptions will be axiom schema instead of rules, and one will be the schema $K(K(\phi) \rightarrow \phi)$. I will discard this schema and show the paradox dissolves.

To see that discarding $K(K(\phi) \rightarrow \phi)$ is philosophically plausible, consider Gödel's Second Incompleteness Theorem: assuming PA is consistent, the arithmetist's knowledge is factive, but she cannot be certain of it. Kritchman and Raz (2010) demonstrated that a similar knowledge paradox, the surprise examination paradox, can be resolved in part by (in essence) altering the definition of *surprise* to say that unsound knowers are surprised by everything. The students in Kritchman and Raz's treatment are sound, but (in the resolution) they interpret their teacher's ambiguous announcement using the modified definition of surprise, acknowleding that they *might* be unsound. The same trick was also used by Halpern and Moses (1986). Factive knowers failing $K(K(\phi) \rightarrow \phi)$ can also be found in the semantics of *impossible-worlds structures* (as defined, for example, by Duc (2001 pp. 21-22)).

But this justification is limited to particular concrete knowers inside closed systems, about whom we possess meta-knowledge. What about our own knowledge, or the total idealized knowledge of mankind? Surely that knowledge is factive (if the barest foundations of science are) and what's more, we know it, or at least we presuppose it in our struggle to define knowledge at all. The situation is similar to how non-paradoxical third-party Moore's paradoxes ("He's factive and he doesn't know it") become paradoxical when there's only one voice ("We're factive and we don't know it"). In short, this paper mainly applies to specific, formal knowers. This is unsatisfying since the most profound questions in the Fitch's Paradox debate are directly concerned with the general meta-knowledge.

Still I can offer one desperate throe in the face of the above. We can speak of *limit knowledge*, saying that a fact is *limit known* if it is eventually always known and always true after some point. Mathematics periodically suffers errors which spread far enough to be considered part of the mathematical knowledge. There may be cofinally many moments when we are presently unfactive, so the statement that we are factive might not be limit knowledge. But all errors are eventually corrected, no error becomes limit knowledge, and limit knowledge is factive.

2. The Paradox

Work in a language $\mathscr{L}$, propositional[1] except for exactly two new unary modal operators K and P. I will give an entirely axiomatic version of the Fitch assumptions, but some machinery is needed. Call a formula *propositional* if it has the form $\rho \wedge \sigma$, $\rho \vee \sigma$, $\rho \to \sigma$, $\rho \leftrightarrow \sigma$, or $\neg\rho$. Now, call a formula *valid* if it can be proved by classic propositional logic treating non-propositional formulas as atoms (an idea inspired by Carlson (2000)). That is, ϕ is valid if and only if it is True according to every truth assignment to the subformulas of ϕ which respects $\wedge$, $\vee$, $\to$, $\leftrightarrow$ and $\neg$.

Thus $K(\phi) \to K(\phi)$ is valid, as is $P(\phi) \vee \neg P(\phi)$, but $K(K(\phi) \to K(\phi))$ is not. Any valid formula holds in any model where the semantics of propositional connectives are classical. Whenever Σ is a set of axioms, write $\Sigma \models \phi$ to mean ϕ can be proved from Σ propositionally, treating non-propositional formulas as atoms.

Our Fitch assumptions are the following set S of axioms:

[1] I originally wanted to publish a first-order treatment, but there are a lot of new technical details. It was H. Friedman who suggested the switch to propositional, which is more in alignment with existing work on Fitch's Paradox anyway.

- E1: $K(\phi)$ whenever ϕ is valid.
- E2: $K(\phi \to \psi) \to K(\phi) \to K(\psi)$.
- E3: $K(\phi) \to \phi$.
- E-Closure: $K(\phi)$ whenever ϕ is an instance of E1, E2, or E3.
- C: $\phi \to P(K(\phi))$.
- D: $K(\neg\phi) \to \neg P(\phi)$.

The names of E1-E3 were chosen to match names used by Carlson (2000 pp. 56). I want to point out we've added no new assumptions not already implied (in classic logic) by Salerno's (2010) more familiar assumptions. For example, to prove E2, assume $K(\phi \to \psi)$ and $K(\phi)$; then by T, $\phi \to \psi$ and ϕ; by modus ponens, ψ; by Fitch's Paradox, $\psi \to K(\psi)$; and by modus ponens, $K(\psi)$. We have simply broken some component pieces off of an inference rule.

Finally we show $S \models \phi \to K(\phi)$. This is a modification of the classical Church-Fitch argument. The original argument was published by Fitch (1963) after it was communicated to him by an anonymous referee who we now know was Alonzo Church. In the following argument (except line 8), we work in propositional logic, treating non-propositional formulas as atoms.

1. Assume $K(\phi \wedge \neg K(\phi))$.
2. By E1, $K((\phi \wedge \neg K(\phi)) \to \phi)$.
3. By E2, $K(\phi \wedge \neg K(\phi)) \to K(\phi)$.
4. By 1, 3, Modus Ponens, $K(\phi)$.
5. An identical argument shows $K(\neg K(\phi))$.
6. By E3, $\neg K(\phi)$.
7. By 4, 6, Contradiction, Discharge 1 and conclude $\neg K(\phi \wedge \neg K(\phi))$.
8. I've proved $\neg K(\phi \wedge \neg K(\phi))$ propositionally from finitely many instances of E1-E3. There are finitely many axioms $\phi_1, ..., \phi_n$ from E1-E3 such that $\phi_1 \to \cdots \to \phi_n \to \neg K(\phi \wedge \neg K(\phi))$ is valid.
9. By E1, $K(\phi_1 \to \cdots \to \phi_n \to \neg K(\phi \wedge \neg K(\phi)))$.
10. By E2 repeatedly, $K(\phi_1) \to \cdots \to K(\phi_n) \to K(\neg K(\phi \wedge \neg K(\phi)))$.
11. We have $K(\phi_1)$, ..., $K(\phi_n)$ by E-Closure.

12. By 10, 11, Modus Ponens, $K(\neg K(\phi \wedge \neg K(\phi)))$.

13. By D, $\neg P(K(\phi \wedge \neg K(\phi)))$.

14. Assume $\phi \wedge \neg K(\phi)$.

15. By C, $P(K(\phi \wedge \neg K(\phi)))$.

16. By 13, 15, Contradiction, Discharge 14 and conclude $\neg(\phi \wedge \neg K(\phi))$.

17. By elementary logic, $\phi \rightarrow K(\phi)$.

Lines 7-12 illustrate how the axioms in S perform work traditionally done by rules of inference. If we strengthened E-closure to also include $K(\phi)$ whenever ϕ is an instance of C, D, or (recursively) E-closure, the resulting system would enjoy the full rule of necessitation $\phi / K(\phi)$.

One feature of this axiomatic version of Fitch's Paradox is that in principle we can use it to write Fitch's Paradox as a *single* logical tautology schema. This can be done by reverse-engineering lines 1-17 above to obtain a tautology of the form $\Psi \rightarrow \phi \rightarrow K(\phi)$, where Ψ is a giant conjunction of specific axioms (depending uniformly on ϕ) of S. Since the conclusion is $K(\phi)$ and one of the hypotheses is $K(K(\neg K(\phi)) \rightarrow \neg K(\phi))$, this is almost a kind of weak propositional version of Löb's Theorem.

3. Avoiding Paradox by Weakening E-Closure

Let S' be the following set of axioms:

- E1: $K(\phi)$ whenever ϕ is valid.
- E2: $K(\phi \rightarrow \psi) \rightarrow K(\phi) \rightarrow K(\psi)$.
- Weak E-Closure: $K(\phi)$ whenever ϕ is an instance of E1 or E2.
- E3: $K(\phi) \rightarrow \phi$.
- C: $\phi \rightarrow P(K(\phi))$.
- D: $K(\neg\phi) \rightarrow \neg P(\phi)$.

We will show these axioms do not imply Fitch's Paradox: resolving it, in the sense of Chow (1998).

Theorem: Let $\mathscr{L}$ be the language, propositional except for two new unary connectives K and P, with a single atom q. Then $S' \not\models q \rightarrow K(q)$.

Proof: Let Σ be the following set of $\mathscr{L}$-axioms:

– E1: $K(\phi)$ whenever ϕ is valid.

– E2: $K(\phi \rightarrow \psi) \rightarrow K(\phi) \rightarrow K(\psi)$.

– Weak E-Closure: $K(\phi)$ whenever ϕ is an instance of E1 or E2.

For each $i \in \{1, 2\}$, let $\mathscr{N}_i$ be the $\mathscr{L}$-model defined as follows:

– $\mathscr{N}_i \models q$ iff $i = 1$.

– $\mathscr{N}_i \models K(\phi)$ iff $\Sigma \models \phi$.

– $\mathscr{N}_i \models P(\phi)$ iff $\Sigma \not\models \neg\phi$.

– $\mathscr{N}_i$ treats propositional sentences inductively in the classical way.

I will show both $\mathscr{N}_i \models S'$. I claim this is enough to prove the theorem: evidently $\mathscr{N}_1$ and $\mathscr{N}_2$ interpret K identically, so if $\mathscr{N}_1 \models K(q)$ then so does $\mathscr{N}_2$, which would contradict that $\mathscr{N}_2 \not\models q$ and $\mathscr{N}_2$ models $K(q) \rightarrow q$. In the following Claims, $\mathscr{N}$ shall stand for either $\mathscr{N}_1$ or $\mathscr{N}_2$ interchangeably.

Claim 1: $\mathscr{N} \models K(\phi)$ if ϕ is valid.

Since ϕ follows propositionally from $\emptyset$, it certainly follows from Σ.

Claim 2: $\mathscr{N} \models K(\phi \rightarrow \psi) \rightarrow K(\phi) \rightarrow K(\psi)$.

Assume $\mathscr{N} \models K(\phi \rightarrow \psi)$ and $\mathscr{N} \models K(\phi)$. Then $\Sigma \models \phi \rightarrow \psi$ and $\Sigma \models \phi$. By Modus Ponens, $\Sigma \models \psi$. So $\mathscr{N} \models K(\psi)$.

Claim 3: $\mathscr{N} \models K(\phi)$ whenever ϕ is an instance of E1 or E2.

If ϕ is an instance of E1 or E2, then $\phi \in \Sigma$, so $\Sigma \models \phi$, so $\mathscr{N} \models K(\phi)$.

Claim 4: $\mathscr{N} \models K(\phi) \rightarrow \phi$.

Assume $\mathscr{N} \models K(\phi)$. Then $\Sigma \models \phi$. By Claims 1-3, $\mathscr{N} \models \Sigma$. Thus $\mathscr{N} \models \phi$.

Claim 5: $\mathscr{N} \models \phi \rightarrow P(K(\phi))$.

In fact, $\mathscr{N} \models P(K(\phi))$, a much stronger fact. To see this, define a new $\mathscr{L}$-model $\mathscr{M}$ as follows: $\mathscr{M} \models q$, $\mathscr{M} \models K(\psi)$ always, $\mathscr{M} \models P(\psi)$ never, and $\mathscr{M}$ treats propositional sentences inductively in the classic way. Then $\mathscr{M}$ trivially satisfies Σ. Since $\mathscr{M} \not\models \neg K(\phi)$, this shows $\Sigma \not\models \neg K(\phi)$. Therefore, $\mathscr{N} \models P(K(\phi))$.

Claim 6: $\mathscr{N} \models K(\neg\phi) \rightarrow \neg P(\phi)$.

Assume $\mathscr{N} \models K(\neg\phi)$. Then $\Sigma \models \neg\phi$. Therefore, it is *not* the case that $\Sigma \not\models \neg\phi$. Thus $\mathscr{N} \not\models P(\phi)$. Thus $\mathscr{N} \models \neg P(\phi)$.

By the above claims, $\mathscr{N} \models S'$, as desired. ∎

This resolution preserves the knowability thesis, that all truths are knowable, which I hope will please the verificationists, though I might have gone too far. In proving Claim 5, we actually showed that in this particular model, *everything* is knowable, whether true or false.

References

– Carlson, T.J. (2000) Knowledge, machines, and the consistency of Reinhardt's strong mechanistic thesis. *Annals of Pure and Applied Logic, 105*, 51-82.

– Chow, T.Y. (1998) The surprise examination or unexpected hanging paradox. *The American Mathematical Monthly 105*, 41-51.

– Duc, H.N. (2001) Resource-Bounded Reasoning about Knowledge. Ph.D. thesis, University of Leipzig.

– Fitch, F. (1963) A logical analysis of some value concepts. *The Journal of Symbolic Logic, 28*, 135-142. (Reprinted in Salerno (Ed.), 2009.)

– Friedman, H. (2011). Personal correspondence.

– Halpern, J., & Moses, Y. (1986) Taken by surprise: The paradox of the surprise test revisited. *Journal of Philosophical Logic 15*, 281-304.

– Kritchman, S. & Raz, R. (2010) The Surprise Examination Paradox and the Second Incompleteness Theorem. *Notices of the American Mathematical Society 57*, 1454-1458.

– Salerno, J. (Ed.). (2009) *New Essays on the Knowability Paradox*. Oxford: Oxford University Press.

– Salerno, J. (2010) Introduction to Knowability and Beyond. This Journal, 173:1-8.