

The Surprise Paradox

Joachim Breitner
Programming Paradigms Group
Karlsruhe Institute for Technology
breitner@kit.edu

23. Februar 2021

Zusammenfassung

In 1964, Fitch showed that the paradox of the surprise hanging can be resolved by showing that the judge's verdict is inconsistent. His formalization builds on Gödel's coding of provability.

In this theory, we reproduce his proof in Isabelle, building on Paulson's formalisation of Gödel's incompleteness theorems.

Inhaltsverzeichnis

1 Excluded or	2
2 Formulas with variables	2
3 Fitch's proof	4
4 Substitution, quoting and V-quoting	5

```
theory Surprise-Paradox  
imports  
  Incompleteness.Goedel-I  
  Incompleteness.Pseudo-Coding  
begin
```

The Surprise Paradox comes in a few variations, one being the following:

A judge sentences a felon to death by hanging, to be executed at noon the next week, Monday to Friday. As an extra punishment, the judge does not disclose the day of the hanging and promises the felon that it will come at a surprise.

The felon, probably a logician, then concludes that he cannot be hanged on Friday, as by then it would not longer be a surprise. Using this fact and similar reasoning, he cannot be hanged on

Thursday, and so on. He reaches the conclusion that he cannot be hanged at all, and contently returns to his cell.

Wednesday, at noon, the hangman comes to the very surprised felon, and executes him.

Obviously, something is wrong here: Does the felon reason wrongly? It looks about right. Or is the judge lying? But his prediction became true!

It is an interesting exercise to try to phrase the Surprise Paradox in a rigorous manner, and see this might clarify things.

In 1964, Frederic Fitch suggested a formulation that refines the notion of “surprise” as “cannot be proven from the given assumptions” [1]. To formulate that, we need propositions that reference their own provability, so just as Fitch builds on Gödel’s work, we build on Paulson’s formalisation of Gödel’s incompleteness theorems in Isabelle [2].

1 Excluded or

Although the proof goes through with regular disjunction, Fitch phrases the judge’s proposition using exclusive or, so we add syntax for that.

abbreviation $Xor :: fm \Rightarrow fm \Rightarrow fm$ (**infix** *XOR* 120)
where $Xor\ A\ B \equiv (A\ OR\ B)\ AND\ ((Neg\ A)\ OR\ (Neg\ B))$

2 Formulas with variables

In Paulson’s formalisation of terms and formulas, only terms carry variables. This is sufficient for his purposes, because the proposition that is being diagonalised needs itself as a parameter to $PfP::tm \Rightarrow fm$, which does take a term (which happens to be a quoted formula).

In order to stay close to Fitch, we need the diagonalised proposition to occur deeper in a quotation of a few logical conjunctions. Therefore, we build a small theory of formulas with variables (“holed” formulas). These support substituting a formula for a variable, this substitution commutes with quotation, and closed holed formulas can be converted to regular formulas.

In our application, we do not need holes under an quantifier, which greatly simplifies things here. In particular, we can use **datatype** and **fun**.

datatype $hfm =$
 $HVar\ name$
 $| HFm\ fm$
 $| HDisj\ hfm\ hfm$ (**infixr** *HOR* 130)
 $| HNeg\ hfm$

abbreviation $HImp :: hfm \Rightarrow hfm \Rightarrow hfm$ (**infixr** *HIMP* 125)

where $HImp\ A\ B \equiv HDisj\ (HNeg\ A)\ B$

definition $HConj :: hfm \Rightarrow hfm \Rightarrow hfm$ (**infixr** *HAND* 135)
where $HConj\ A\ B \equiv HNeg\ (HDisj\ (HNeg\ A)\ (HNeg\ B))$

abbreviation $HXor :: hfm \Rightarrow hfm \Rightarrow hfm$ (**infix** *HXOR* 120)
where $HXor\ A\ B \equiv (A\ HOR\ B)\ HAND\ (HNeg\ A\ HOR\ HNeg\ B)$

fun $subst\text{-}hfm :: hfm \Rightarrow name \Rightarrow fm \Rightarrow hfm$ ($\text{'(-::=-)'} [1000, 0, 0] 200$)
where
 $(HVar\ name)(i::=x) = (if\ i = name\ then\ HFm\ x\ else\ HVar\ name)$
 $| (HDisj\ A\ B)(i::=x) = HDisj\ (A(i::=x))\ (B(i::=x))$
 $| (HNeg\ A)(i::=x) = HNeg\ (A(i::=x))$
 $| (HFm\ A)(i::=x) = HFm\ A$

lemma $subst\text{-}hfm\text{-}Conj[simp]$:
 $(HConj\ A\ B)(i::=x) = HConj\ (A(i::=x))\ (B(i::=x))$
 $\langle proof \rangle$

instantiation $hfm :: quot$
begin
fun $quot\text{-}hfm :: hfm \Rightarrow tm$
where
 $quot\text{-}hfm\ (HVar\ name) = (Var\ name)$
 $| quot\text{-}hfm\ (HFm\ A) = \langle A \rangle$
 $| quot\text{-}hfm\ (HDisj\ A\ B) = HPair\ (HTuple\ 3)\ (HPair\ (quot\text{-}hfm\ A)\ (quot\text{-}hfm\ B))$
 $| quot\text{-}hfm\ (HNeg\ A) = HPair\ (HTuple\ 4)\ (quot\text{-}hfm\ A)$

instance $\langle proof \rangle$
end

lemma $subst\text{-}quot\text{-}hfm[simp]$: $subst\ i\ \langle P \rangle\ \langle A \rangle = \langle A(i::=P) \rangle$
 $\langle proof \rangle$

fun $hfm\text{-}to\text{-}fm :: hfm \Rightarrow fm$
where
 $hfm\text{-}to\text{-}fm\ (HVar\ name) = undefined$
 $| hfm\text{-}to\text{-}fm\ (HFm\ A) = A$
 $| hfm\text{-}to\text{-}fm\ (HDisj\ A\ B) = Disj\ (hfm\text{-}to\text{-}fm\ A)\ (hfm\text{-}to\text{-}fm\ B)$
 $| hfm\text{-}to\text{-}fm\ (HNeg\ A) = Neg\ (hfm\text{-}to\text{-}fm\ A)$

lemma $hfm\text{-}to\text{-}fm\text{-}Conj[simp]$:
 $hfm\text{-}to\text{-}fm\ (HConj\ A\ B) = Conj\ (hfm\text{-}to\text{-}fm\ A)\ (hfm\text{-}to\text{-}fm\ B)$
 $\langle proof \rangle$

fun $closed\text{-}hfm :: hfm \Rightarrow bool$
where
 $closed\text{-}hfm\ (HVar\ name) \longleftrightarrow False$
 $| closed\text{-}hfm\ (HFm\ A) \longleftrightarrow True$

| *closed-hfm* (*HDisj* *A B*) \longleftrightarrow *closed-hfm* *A* \wedge *closed-hfm* *B*
| *closed-hfm* (*HNeg* *A*) \longleftrightarrow *closed-hfm* *A*

lemma *closed-hfm-Conj*[*simp*]:
closed-hfm (*HConj* *A B*) \longleftrightarrow *closed-hfm* *A* \wedge *closed-hfm* *B*
⟨*proof*⟩

lemma *quot-closed-hfm*[*simp*]: *closed-hfm* *A* \implies «*A*» = «*hfm-to-fm* *A*»
⟨*proof*⟩

declare *quot-hfm.simps*[*simp del*]

3 Fitch's proof

For simplicity, Fitch (and we) restrict the week to two days. Propositions Q_1 and Q_2 represent the propositions that the hanging occurs on the first resp. the second day, but these can obviously be any propositions.

context

fixes $Q_1 :: fm$ **and** $Q_2 :: fm$
assumes *Q-closed*: *supp* $Q_1 = \{\}$ *supp* $Q_2 = \{\}$

begin

In order to define the judge's proposition, which is self-referential, we apply the usual trick of defining a proposition with a variable, and then using Gödel's diagonalisation lemma.

definition *H* :: *fm* **where**

$H = Q_1 \text{ AND } \text{Neg} (\text{PfP} \langle \text{HVar } X0 \text{ HIMP } \text{HFm } Q_1 \rangle) \text{ XOR}$
 $Q_2 \text{ AND } \text{Neg} (\text{PfP} \langle \text{HVar } X0 \text{ HAND } \text{HNeg} (\text{HFm } Q_1) \text{ HIMP} (\text{HFm } Q_2) \rangle)$

definition *P* **where** $P = (\text{SOME } P. \{\} \vdash P \text{ IFF } H(X0 ::= \langle P \rangle))$

lemma *P'*: $\{\} \vdash P \text{ IFF } H(X0 ::= \langle P \rangle)$
⟨*proof*⟩

From now on, the lemmas are named after their number in Fitch's paper, and correspond to his statements pleasingly closely.

lemma 7: $\{\} \vdash P \text{ IFF}$
 $(Q_1 \text{ AND } \text{Neg} (\text{PfP} \langle P \text{ IMP } Q_1 \rangle) \text{ XOR}$
 $Q_2 \text{ AND } \text{Neg} (\text{PfP} \langle P \text{ AND } \text{Neg } Q_1 \text{ IMP } Q_2 \rangle))$
⟨*proof*⟩

lemmas 7-E = 7[*THEN thin0*, *THEN Iff-MP-left'*, *OF Conj-E*, *OF thin2*]

lemmas *propositional-calculus* =

AssumeH *Neg-I* *Imp-I* *Conj-E* *Disj-E* *ExFalso*[*OF Neg-E*]
ExFalso[*OF rotate2*, *OF Neg-E*] *ExFalso*[*OF rotate3*, *OF Neg-E*]

lemma 8: $\{\} \vdash (P \text{ AND } \text{Neg } Q_1) \text{ IMP } Q_2$

$\langle proof \rangle$

lemma 10: $\{\} \vdash PfP \langle (P \text{ AND } Neg Q_1) \text{ IMP } Q_2 \rangle$

$\langle proof \rangle$

lemmas 10-I = 10[THEN thin0]

lemma 11: $\{\} \vdash P \text{ IMP } Q_1$

$\langle proof \rangle$

lemma 12: $\{\} \vdash PfP \langle P \text{ IMP } Q_1 \rangle$

$\langle proof \rangle$

lemmas 12-I = 12[THEN thin0]

lemma 13: $\{\} \vdash Neg P$

$\langle proof \rangle$

end

4 Substitution, quoting and V-quoting

In the end, we did not need the lemma at the end of this section, but it may be useful to others.

lemma trans-tm-forgets: $atom \text{ ' set is } \#* t \implies trans\text{-tm is } t = trans\text{-tm } [] t$

$\langle proof \rangle$

lemma vquot-dbtm-fresh: $atom \text{ ' } V \#* t \implies vquot\text{-dbtm } V t = quot\text{-dbtm } t$

$\langle proof \rangle$

lemma subst-vquot-dbtm-trans-tm[simp]:

$atom \text{ ' } \# is \implies atom \text{ ' set is } \#* t \implies$
 $subst \text{ ' } \langle t \rangle (vquot\text{-dbtm } \{i\} (trans\text{-tm is } t')) =$
 $quot\text{-dbtm } (trans\text{-tm is } (subst \text{ ' } t t'))$

$\langle proof \rangle$

lemma subst-vquot-dbtm-trans-fm[simp]:

$atom \text{ ' } \# is \implies atom \text{ ' set is } \#* t \implies$
 $subst \text{ ' } \langle t \rangle (vquot\text{-dbfm } \{i\} (trans\text{-fm is } A)) =$
 $quot\text{-dbfm } (trans\text{-fm is } (subst\text{-fm } A \text{ ' } t))$

$\langle proof \rangle$

lemma subst-vquot[simp]:

$subst \text{ ' } \langle t \rangle [A] \{i\} = \langle A(i ::= t) \rangle$

$\langle proof \rangle$

end

Literatur

- [1] F. B. Fitch. A goedelized formulation of the prediction paradox. *American Philosophical Quarterly*, 1(2):161–164, 1964.
- [2] L. C. Paulson. Gödel’s incompleteness theorems. *Archive of Formal Proofs*, Nov. 2013. <http://isa-afp.org/entries/Incompleteness.shtml>, Formal proof development.